Landuo Dou and Xiaodong Dou*

Towards Just AI: Challenges and Solution Framework for Algorithmic Discrimination in Judicial System

https://doi.org/10.1515/ijdlg-2024-0020
Received November 11, 2024; accepted February 9, 2025; published online March 5, 2025

Abstract: With the rapid development of technologies like big data, artificial intelligence (herein after AI), and blockchain, society is ushering into a new era of digital civilization. However, the same algorithms that assist in efficient decision-making for human society may also introduce issues of algorithmic discrimination. Therefore, this paper focuses on the judicial domain to deeply explore potential instances of algorithmic discrimination in AI. It identifies three key dimensions of algorithmic discrimination risks in the judicial AI domain: the ambiguity of algorithm usage boundaries, the diversity in discriminatory outcomes, and the injustice within the algorithmic environment. Building upon this, the author examines global AI governance landscapes to extract corresponding governance strategies and practical insights. Finally, a systematic regulatory approach for addressing algorithmic discrimination in judicial AI is proposed, unfolding across three levels and nine aspects: making algorithmic limitations explicit, diversifying algorithmic regulations, and justifying the algorithmic environment. This framework aims to contribute to a more reasonable, systematic, and just governance of algorithms in judicial AI.

Keywords: judicial artificial intelligence; algorithmic discrimination; digital justice; legal regulation

1 Introduction

The widespread adoption of modern digital tools such as the Internet, cloud computing, big data, and artificial intelligence (AI) is ushering in a transformative era for society. This shift is not only altering productivity and production relations

***Corresponding author: Xiaodong Dou**, Guanghua Law School, Zhejiang University, Hangzhou, China; and Institute of New Era Fengqiao Experience, Zhejiang University, Hangzhou, China, E-mail: douxd@zju.edu.cn
Landuo Dou, Guanghua Law School, Zhejiang University, Hangzhou, China; and Law and Technology Institute, Renmin University, Beijing, China, E-mail: landuo_dou@zju.edu.cn

but also reshaping both the economic base and the superstructure. As human activities and social interactions increasingly become digitalized, a new digital identity is emerging (Song and Ma 2022). This societal transition signals the dawn of the "Digital Civilization Era," where the convergence of digitization and intelligence is reshaping many domains, including judicial adjudication.

Judicial AI has its intellectual roots in the broader development of "legal science," a concept introduced by German mathematician and philosopher Gottfried Wilhelm Leibniz in 1663 (Gottfried 2013). In his work Philosophical Problems in the Law: Logical Puzzles, Leibniz advocated for the use of computational methods to address legal and philosophical problems, laying the groundwork for what would later be recognized as legal science. This early integration of scientific methods into legal reasoning was further supported by prominent legal thinkers such as Justice Oliver Wendell Holmes, who argued that a truly ideal legal system must be grounded in science (Holmes 1946), and Judge Benjamin Nathan Cardozo, who proposed using mathematical formulas to derive "justice" (Cardozo 1928). These theoretical underpinnings played a crucial role in the eventual empirical and data-driven transformations that would characterize contemporary legal practice.

The integration of AI into judicial processes has the potential to revolutionize the legal landscape, offering efficiencies in case processing, decision-making, and resource allocation. However, as AI systems – particularly those used in judicial decision-making – become more prevalent, significant concerns have emerged about the possibility of algorithmic discrimination. Algorithmic discrimination, or the systematic bias in AI systems that leads to unfair or disproportionate outcomes for certain groups, poses a critical challenge for the equitable application of justice in AI-enhanced legal systems. Judicial AI systems, which include tools for predictive policing, risk assessments, and sentencing recommendations, have been subject to growing scrutiny regarding the biases embedded in their underlying algorithms (Angwin, Larson, and Mattu 2022).

Algorithmic discrimination in judicial AI can manifest in numerous ways, such as racial or gender bias in risk prediction tools (e.g. COMPAS:Correctional Offender Management Profiling for Alternative Sanctions) or biased sentencing recommendations. These systems, which rely on historical data, may perpetuate existing societal inequalities by reinforcing patterns of discrimination embedded in past decisions. Scholars have increasingly argued that AI systems, when not carefully regulated, can inadvertently exacerbate social inequities and disproportionately harm marginalized communities (Eubanks 2018; Noble 2018). For example, the COMPAS risk assessment algorithm, which was widely used in the United States for parole decisions, has been found to overestimate the risk of reoffending among Black defendants compared to white defendants, raising significant concerns about racial fairness in AI-driven judicial processes (Angwin, Larson, and Mattu 2022).

Released on Netflix, the leading platform known for its algorithm-driven content, The Social Dilemma (TSD) vividly illustrates how recent breakthroughs in AI and Machine Learning Algorithms (MLAs) have transformed these technologies into a new form of post-digital, semi-cognitive power (Elyamany 2024). TSD starkly highlights the unprecedented influence AI holds in shaping public opinion, reinforcing biases, and even influencing democratic processes. This power is no less evident in the judicial sphere, where algorithmic decision-making, when unchecked, can produce biased and inequitable outcomes, raising critical questions about accountability and control over AI-driven systems in legal contexts.

In the judicial context, algorithmic discrimination manifests in three critical dimensions: the ambiguity of algorithm usage boundaries, the diversity of discriminatory outcomes, and the inherent injustice within the algorithmic environment. First, the ambiguity surrounding the scope and application of judicial AI raises questions about the ethical and legal limits of algorithmic decision-making. This uncertainty can lead to inconsistent and opaque practices that undermine public trust in the judicial system (Sandvig 2022; Zarsky 2021). Second, the diversity in discriminatory outcomes – ranging from racial and gender biases to socioeconomic prejudice – illustrates the complex and multifaceted nature of algorithmic harm. For instance, predictive algorithms like COMPAS have been shown to disproportionately label Black defendants as high-risk, while underestimating the recidivism risk of white defendants. Finally, the algorithmic environment itself often lacks the safeguards necessary to ensure fairness and accountability, with AI systems being prone to entrenching existing inequalities due to flawed or biased training data (O'Neil 2016).

Building on these challenges, this paper proposes a systematic regulatory framework designed to mitigate algorithmic discrimination in judicial AI. The framework unfolds across three levels: (1) making the limitations of judicial AI algorithms explicit, (2) diversifying regulatory approaches to encompass the complex nature of algorithmic bias, and (3) justifying the algorithmic environment to ensure accountability, transparency, and fairness. By addressing these three core areas, the framework aims to provide a more reasonable, systematic, and just approach to governing AI in the judicial domain. Ultimately, the goal is to provide regulatory guidance to ensure that AI can better support human tasks in a more just, safe, and ethical manner, contributing to the further development of digital civilization.

2 The Implications of Algorithmic Discrimination in Judical AI

2.1 The Concept of Algorithmic Discrimination

The Oxford Living Dictionary defines algorithms as “processes or sets of rules to be followed in calculations or other problem-solving operations, especially by a computer.”

In computer science, an algorithm is typically understood as a formalized procedure that takes an input, processes it, and yields an output after a finite number of steps. Each step is precisely defined, unambiguous, and effective. As Turing (1936) notes, an algorithm is a sequence of well-defined instructions for carrying out a task in a finite amount of time (Turing 1936). Generally speaking, algorithms generate instructions through code and effectively fulfill the anticipated human needs. Therefore, algorithms fundamentally embody human thought, serving as a means and method to address practical problems through the utilization of data, logic, and code, among other forms. Algorithms can typically be categorized into two modes: fully automated and semi-automated. Fully automated algorithms provide direct answers to problems, influencing stakeholders directly, while semi-automated algorithms assist human decision-making by offering alternative options or bases for decision outcomes, facilitating subsequent human processing. In the current era of rapid digital civilization development, algorithmic discrimination based on the inherent data and technical attributes of algorithms is gradually emerging. This phenomenon may lead to unjust treatment of specific individuals or groups. Algorithmic discrimination is a form of discrimination arising from machine-automated decision-making. Compared to traditional discrimination, algorithmic discrimination possesses more complex mechanisms of occurrence, with various processes such as problem design, data selection, data filtering, and model design potentially contributing to discriminatory outcomes. Furthermore, due to the widespread application of digital technology, if algorithmic discrimination occurs, it may have broader and deeper-ranging impacts.

The conceptualization of algorithmic discrimination has evolved significantly over the past few decades, as scholars from diverse fields – law, computer science, sociology, and ethics – have engaged with its implications in different social, legal, and technological contexts. Initially, the idea of algorithmic discrimination was largely influenced by statistical discrimination theories from economics and law, which focus on how statistical proxies, such as race, gender, and are used by algorithms to predict outcomes. This early interpretation of algorithmic discrimination emerged from concerns about disparate impact in algorithmic decision-making. For example, Early interpretations, like those advanced by Kenneth Arrow, viewed discrimination as the use of biased or incomplete information to make decisions that lead to unequal outcomes for different social groups. In this context, statistical discrimination occurs when algorithms rely on historical data that reflects past inequities, such as disproportionately high arrest rates for minority groups, leading to biased predictions (Arrow 1971).

In the 2010s, scholars began to shift the focus from purely statistical discrimination to broader ethical and sociotechnical considerations of algorithmic discrimination. This transition was fueled by the rapid deployment of machine learning in sensitive areas such as criminal justice, hiring, and finance, which raised concerns about the fairness and accountability of algorithms. In this period, a key turning point came with the publication of works like Cathy O'Neil's "Weapons of Math Destruction" (O'Neil 2016), which argued that algorithms, when used irresponsibly, could harm vulnerable populations by reinforcing existing biases in data. O'Neil emphasized how algorithms disproportionately affect marginalized communities by disproportionately penalizing them through opaque, unregulated systems. Besides, critical theorists, particularly Shoshana Zuboff and Safiya Noble, began to frame algorithmic discrimination in the context of surveillance capitalism and digital colonialism. They argued that algorithms are not neutral tools but are embedded in larger systems of power and control. These scholars contend that algorithmic discrimination is not merely a product of biased data, but also a consequence of the unequal distribution of power in the design and deployment of AI technologies, which tend to benefit dominant groups and institutions (Nobel 2018; Zuboff 2023).

With the rapid development of AIGC, AI is playing an increasingly important role around the world. In light of global concerns around AI's role in perpetuating inequality, discourse has seen a push for international regulations and standards to govern AI and combat discrimination. This view stresses that algorithmic discrimination is a global issue, with systemic bias affecting populations worldwide. As AI systems are increasingly used in transnational contexts, scholars call for global collaboration and multilateral governance frameworks. For example, Floridi's work on AI ethics emphasizes the need for global cooperation on AI governance. He argues that there is a need for international agreements on AI to ensure that algorithms respect human rights and ethical standards globally. In particular, He advocates for creating global digital rights frameworks to protect citizens against algorithmic discrimination and exploitation (Floridi 2023).

2.2 Key Characteristics of Algorithmic Discrimination in Judicial AI

2.2.1 The Complexity of Algorithmic Discrimination in Legal Decision-Making

Firstly, the characteristic of complexity is evident in the existence of algorithmic discrimination in the field of judicial AI. In comparison to traditional forms of discrimination, identifying the sources and reasons for the occurrence of algorithmic discrimination is a decisive challenge. For instance, in traditional contexts, whether it be the requirements for employee recruitment in corporate processes or policies enacted by governments targeting specific groups (Ding 2014), the perpetrators of discrimination and the reasons for its occurrence can be readily identified.

However, in the case of algorithmic discrimination, despite its close correlation with the entire design process and human behavior, the sources and mechanisms of discrimination are not explicitly known. On one hand, algorithms have exceptionally broad data sources, ranging from government open data, official documents, to judicial institutions having access to original case files and legal texts. In addition to regulated and explicit data, various types of trace records and open-source materials available on the internet can be incorporated into the training set. When these data are reorganized and processed, it becomes challenging to attribute discrimination to a specific source. On the other hand, the "black box problem" in machine learning, where the reasoning behind decisions is opaque, further complicates the issue of algorithmic discrimination in the application of judicial AI. For example, COMPAS is a widely used AI software in North America, primarily designed to predict the risk of recidivism among offenders. Developed by a company formerly known as Northpointe, its purpose is to assist judges and parole boards in making informed decisions regarding bail, sentencing, and parole. However, a study revealed that the COMPAS AI crime prediction system used by U.S. courts mistakenly marked Black defendants as twice as likely to reoffend compared to White defendants (Feller et al. 2016). The intricate mechanisms of algorithmic discrimination make the discovery and resolution of discrimination extremely challenging. Without timely adjustments and training, judicial AI systems like COMPAS may perpetuate biases and discrimination, forming a feedback loop that reinforces existing prejudices and inequalities. For instance, if COMPAS consistently overestimates the risk of reoffending for Black defendants, this could lead to harsher sentencing or denial of parole, which in turn could contribute to higher incarceration rates within Black communities. This cycle further exacerbates the development of discrimination, as the system's predictions are influenced by historical and systemic inequities that are already present in the criminal justice system.

2.2.2 The Recurrence of Bias in Judicial AI Systems

Secondly, a characteristic of algorithmic discrimination in the field of judicial AI is its repetitiveness. German sociologist Max Weber once proposed, modern judges are like vending machines, where the public puts in complaints and litigation fees, and what comes out are judgments and reasons copied from the legal code (Weber 2019). The process of judicial adjudication is likened to the workings of a machine, operating through fixed procedures. With the current trend toward intelligent judicial adjudication, especially with the use of AI to recapitulate judicial decisions, the original intention is to alleviate the pressure on the courts, freeing up judicial productivity. Thus, it adheres to the goal of "design once, use infinitely". Admittedly, the program design of AI will be periodically adjusted and corrected based on specific circumstances, but the overall design philosophy and algorithmic framework do not undergo significant changes. Operating under this mechanism, if there is a tendency or possibility of algorithmic discrimination, it can mechanically lead to repetitive occurrences of algorithmic discrimination. As for traditional judicial adjudication, each judge's decision is relatively independent and less likely to influence subsequent rulings. Precedent guides judges, but they can distinguish a current case from previous ones if the facts differ. Judges are not forced to follow an earlier decision if they believe it is not applicable or relevant to the new case. Besides, judges have discretion to interpret the law, which allows them to adapt their decisions to the specifics of each case. Even when a precedent exists, judges can choose not to apply it if they find the facts or legal context significantly different. However, for judicial AI, if algorithmic discrimination is embedded early in the system, it can persistently reoccur in subsequent cases, thereby exerting a long-term influence on the judicial adjudication system.

For instance, India's "Aadhaar" system aims to provide each citizen with a unique identification number through technologies such as fingerprint, iris scanning, and facial recognition (Rao and Nair 2019). Initially designed to enhance the efficiency of government services, such as the distribution of social welfare, pensions, and food subsidies, the Aadhaar system has gradually been applied in the judicial domain. It is used to verify the identities of defendants and, in conjunction with other data (e.g., criminal records, socioeconomic backgrounds), generates risk scores to assist judges in making bail or sentencing decisions. However, the system has exposed severe algorithmic bias issues in practice, particularly discrimination against low-income groups and marginalized communities. The Aadhaar system relies on citizens' biometric data and government databases. However, India's socioeconomic inequalities result in many low-income individuals and marginalized communities lacking high-quality biometric data (e.g., worn fingerprints or failed facial recognition), leading to inaccurate identification by the system. This data bias causes these groups to be erroneously flagged as “high-risk” or “untrustworthy” in judicial decision-making. When the Aadhaar system incorrectly labels a group as “high-risk,” these individuals may be denied bail or subjected to harsher sentencing. Such unjust rulings, in turn, increase their criminal records, further reinforcing the system’s “high-risk” label for them. This feedback loop perpetuates algorithmic discrimination, creating a vicious cycle that is difficult to break.

Therefore, for intelligent algorithms, the decision-making mechanism can undergo adjustments to a certain extent based on specific changes in data, application scenarios, and correction mechanisms. However, if there are insufficient review and oversight mechanisms in the system design, solely relying on the subjective judgment of algorithm designers to determine the existence of algorithmic discrimination and the need for algorithm maintenance, addressing algorithmic discrimination in a timely manner becomes challenging. Such a discriminatory decision-making mechanism is likely to be sustained or even strengthened, contributing to a polarization phenomenon of “the strong becoming stronger and the weak becoming weaker (Sun 2019).”

2.2.3 The Clandestine Nature of Algorithmic Bias in Judicial AI

Finally, a characteristic of algorithmic discrimination in the field of judicial AI is its clandestine nature. In scenarios where judicial AI may give rise to algorithmic discrimination, predicting crimes and assisting in sentencing are common contexts where discrimination may covertly occur. Taking sentencing assistance as an example, during the process of algorithmic decision support, factors such as the severity of the crime, criminal history, and social background may be considered to determine the length of the sentence and whether to grant parole. However, these factors may not be entirely objective and impartial. The algorithm may excessively weigh criminal records without considering the possibility of rehabilitation for the offender. However, the consideration of parole itself is a negative action, meaning that even if the victim is discriminated against, there is no tangible and immediate perception of harm to personal or property rights, only enduring a negative adverse consequence. As Barocas and Selbst (2016) point out, the operation of these systems often lacks transparency, making it challenging for individuals affected by these decisions to understand or challenge them (Barocas and Selbst 2016). This opacity contributes to the clandestine nature of algorithmic discrimination, as the underlying data and model parameters are often proprietary and inaccessible for scrutiny.

On one hand, victims seeking information about cases similar to their own but with different judicial outcomes, to establish the likelihood of discrimination, face significant challenges. With the increasing diversity of elements considered by current AI, case classifications no longer rely on simple factors like gender, race, or ethnicity. Consequently, it becomes difficult for claimants to identify deeper levels of similarity. On the other hand, victims are constrained by the "black box" mechanism of algorithms, unable to access the basis and process of decision-making. They can only assume that the results depend on data and algorithms, following basic mathematical logic. Moreover, the perceived objectivity directly brought about by the algorithm itself often prevents many actual victims from reflecting on the potential for discrimination or being aware of its occurrence. Additionally, for users applying the algorithm, if the hidden processes are utilized for assisting judicial decisions, the "black box" mechanism serves as a shield, allowing them to better conceal the origin of discrimination and escape legal responsibility. In general, these algorithms can reflect and perpetuate inequalities inherent in the criminal justice system (Kleinberg et al. 2018). By relying on such biased inputs, the algorithm can assist in sentencing in a manner that disproportionately affects marginalized groups, despite the apparent neutrality of the tool.

3 The Challenges of Algorithmic Discrimination in Judicial AI

3.1 Ambiguities in the Boundaries of Algorithmic Usage in the Judiciary

3.1.1 Defining the Limits of Algorithmic Application in Legal Decision-Making

AI, as an auxiliary tool in judicial processes, has been bestowed with high expectations for facilitating the determination of facts in criminal cases and assisting judges in rendering judgments. Particularly in the face of a continuous rise in the total number of cases, a compression of individual case trial times, and an increasing workload on judges, judicial AI has progressively played a larger role in the realm of adjudication assistance. However, as cases are propelled forward through data-driven and pattern-based technologies, the judicial authority of legal professionals is inadvertently diminished, and the power of judgment gradually shifts from humans to machines. If this development model persists, a significant transformation will occur in the subject of fact determination in judicial practice: transitioning from traditional legal provisions + judicial rulings to a comprehensive judgment involving data engineers, software designers, and judges. Such a transformation risks disrupting the traditional structures of fact determination and adjudicative power, affecting the fundamental framework of judicial procedures. Some person even sites that as AI systems approach the point of indistinguishability from humans they should be entitled to a status comparable to natural persons (Chesterman 2020). This shift, characterized by the migration of judgment power from legal professionals to algorithmic systems, raises critical questions about the integrity and fairness of the judicial process. The potential for AI to disrupt established structures of fact determination and adjudicative power is not merely theoretical; it is a pressing concern that could fundamentally alter the nature of judicial practice.

Moreover, the rise of AI in judicial systems, while promising increased efficiency and consistency, also necessitates a rigorous examination of the boundaries within which these technologies should operate. The challenge lies in defining these boundaries, as AI tools, if not adequately governed, risk entrenching existing social biases and undermining the principles of justice. In judicial contexts, where the stakes are high and the consequences of biased decisions are severe, it is imperative to establish clear operational limits for AI technologies to prevent the exacerbation of pre-existing inequalities (Cath 2018). The complexity of setting these limits is further compounded by the current legal landscape, which often lacks specificity regarding algorithmic discrimination in judicial applications. For instance, while the EU's AI Act proposes broad rules for "high-risk" systems, the lack of detailed guidelines leaves many aspects open to interpretation, potentially undermining the effectiveness of these regulations in sensitive judicial contexts (Schwemer, Tomada, and Pasini 2021).

3.1.2 The Tension Between Fairness and Justice in Algorithmic Systems

In the judicial system, fairness and justice are two core values that together form the cornerstone of the legal framework. Fairness emphasizes the equality and consistency of legal application, ensuring that all individuals are treated equally before the law. Justice, on the other hand, demands that legal decisions reflect the uniqueness of each case and its social context, ensuring that the verdict aligns with moral and societal standards of fairness. Justice ensures that judicial outcomes resonate with the ethical and social standards of the society they serve, an idea well-expressed in Dworkin's *Law's Empire*, where he contends that law must not merely be a set of rules but should align with principles of justice (Dworkin 1986). However, with the widespread adoption of AI technologies within the judicial system, the tension between fairness and justice has become increasingly apparent. While the introduction of algorithms aims to enhance efficiency and consistency, it also poses potential threats that could undermine the fairness and justice of judicial processes. The application of AI in the judicial system is often intended to enhance the fairness of legal application through data-driven decision-making. The core advantages of algorithms lie in their ability to provide "consistency" and "standardization," thereby reducing the influence of human bias and subjective judgment. Firstly, consistency allows algorithms to deliver uniform judgment recommendations for similar cases by following predefined rules and data models. For instance, in sentencing decisions, algorithms can generate standardized sentencing recommendations based on factors such as crime type, criminal history, and social background, thereby avoiding unfair judgments caused by a judge's personal bias. This consistency helps ensure the equality of legal application, thereby achieving the goal of fairness. Secondly, the standardization of algorithms minimizes subjectivity and arbitrariness in judicial decision-making. For example, in bail decisions, algorithms can provide consistent bail recommendations based on a defendant's recidivism risk score, avoiding unfair outcomes resulting from a judge's personal emotions or biases. This standardized approach enhances the transparency and predictability of judicial decisions, thereby strengthening public trust in the judicial system.

However, despite the technical support provided by algorithmic consistency and standardization for fairness, their mechanical and simplified nature may pose a threat to the realization of justice, thereby exacerbating the tension between fairness and justice in the context of AI in judiciary. This technical focus, according to legal scholars like Richard Posner (Posner 2014), risks neglecting the broader social and ethical obligations that are integral to the administration of justice. Posner's critique of the overreliance on efficiency and cost-benefit analysis in legal contexts is particularly relevant when considering algorithmic systems that might optimize for fairness at the expense of justice, potentially exacerbating inequality. Justice requires that legal decisions reflect the uniqueness of each case, including factors such as the defendant's attitude toward guilt, expressions of remorse, and social impact. The essence of justice lies in ensuring that verdicts not only comply with legal regulations but also reflect the specific circumstances of the case and societal standards of fairness. Each case has its unique factual background and social impact, which may significantly influence the verdict. For instance, in some cases, a defendant's voluntary confession, expressions of remorse, and willingness to compensate the victim may significantly affect the sentencing outcome. However, algorithms typically rely solely on objective data (such as criminal records and socioeconomic background) and may fail to fully capture these subjective factors, potentially leading to verdicts that contradict principles of justice. Furthermore, each individual involved in a case may have specific needs that are difficult to account for in the context of AI-driven judicial processes. For example, in some cases, a defendant may require a lighter sentence due to family responsibilities or social contributions, but algorithms may not fully consider these factors, leading to a neglect of individual justice.

In conclusion, while AI technologies offer promising solutions to enhance the fairness of judicial processes through consistency and standardization, they also introduce challenges that may compromise the pursuit of justice. Balancing the mechanical efficiency of algorithms with the nuanced demands of justice remains a critical challenge in the integration of AI within the judicial system.

3.1.3 The Imbalance Between Formal and Substantive Reasons in Judicial AI

According to the reasons forming the final judicial judgment, common reasons can be categorized into substantive reasons and formal reasons. Formal reasons refer to those reasons grounded in the structure and procedural rules of a system. In the context of law, formal reasons are those that comply strictly with established legal procedures, logical coherence, and legal formalism. It is closely related to the formalism of law, and is placed on consistency, predictability, and adherence to established norms or frameworks, as seen in classical legal theories such as those espoused by Hans Kelsen or Jeremy Bentham. These systems prioritize rule-following over outcomes in terms of justice or fairness (Bentham 1970; Kelsen 2017).

Substantive reasons involve considerations related to the content or outcomes of a decision. These reasons are concerned with the justice, equity, or morality of a decision. In legal contexts, substantive reasons seek to address fairness, justice, and the deeper values of the legal system – things like individual rights, social fairness, and equality. Substantive reasons draw from legal realism, which argues that laws should reflect the real-world outcomes and social needs rather than abstract formal principles. This perspective is closely associated with theorists like Oliver Wendell Holmes Jr., who argued that law must be understood not only by its formal rules but by its impact on people and society (Holmes 1997). Substantive reasons focus on the core issues of the case and the substantive significance of the law, typically involving the purposes, principles, and values of the law. In the application of substantive reasons, the court or decision-making body considers the purpose of the law to ensure fairness, reasonableness, and compliance with social justice in the ruling, relying more on external legal arguments to justify the legitimacy and rationality of legal judgments. Formal reasons, on the other hand, refer to the legal procedures and procedural provisions that the court or decision-making body relies on when reviewing cases or making decisions. Formal reasons concentrate on aspects such as the trial procedure, the legality of evidence, and compliance with statutory procedures. For example, arguing the proposition "cannot kill" in judicial judgments can be interpreted from a substantive perspective as fulfilling the basic requirement of social order and the intrinsic value of the right to life, while from a formal perspective, it can be derived from Criminal Law Code.

In the traditional judicial adjudication process, there is inherent tension and imbalance between formal and substantive reasons. Under the guiding principle of "adjudication according to law", the judicial system emphasizes the legality and fairness of the procedures, focusing on the application of formal reasons, while to some extent neglecting substantive reasons. This imbalance makes it difficult for certain special cases to obtain individual justice within the vast volume of judicial judgments. In the scenario of AI-assisted judicial adjudication, the imbalance between formal and substantive reasons may be exacerbated, and the mechanisms for its occurrence mainly exist in the following two technologies:

Imbalance in the argumentation of formal and substantive reasons caused by the use of automated decision-making tools. In the AI-assisted judicial adjudication process, these computational tools make judgments based on preset algorithms and rules, emphasizing procedural provisions such as data accuracy, the legality of evidence, and the regularity of procedures. While this processing method is suitable for simple and common cases, when applied to the handling of complex cases, especially those with conflicting rights and values, AI may fail due to its difficulty in fully fitting the human social value system and moral thinking. As of the current development of AI, such tools do not possess a comprehensive ability to interpret the purpose, principles, and values of the law, thus potentially being unable to fully consider the fairness and justice of cases. Ethically, the reliance on AI in judicial decision-making may compromise the legitimacy of legal systems by detaching decisions from human values and context. The shift from human judgment to automated processes in law may threaten to erode accountability and undermine public trust in the justice system (Pasquale 2015).

Potential risks in the imbalance of formal and substantive reasons induced by big data analysis. AI-assisted judicial adjudication systems need to learn a large amount of public social data, government open data, and existing case analyses during the training phase, based on big data analysis technology, involving unsupervised, semi-supervised, and supervised learning. On the one hand, this reliance on data and statistical information may lead the court or decision-making body to overlook the specific circumstances of individual cases and the specific application of legal principles, thereby exacerbating the traditional imbalance between formal and substantive reasons. On the other hand, due to the extensive nature of big data, careful discrimination between true and false information is challenging, and AI produced based on massive data training may exacerbate inequality and increase the likelihood of errors, thereby contradicting the original intention of judicial adjudication. Therefore, some scholar argues that rather than replacing judges, AI should be seen as a tool that enhances their ability to make more informed and fair decisions, with a greater focus on substantive justice (Binns 2018).

3.2 The Multifaceted Nature of Algorithmic Discrimination Risks

3.2.1 Algorithmic Discrimination Risk in the Dimension of Problem Formulation

The objective of this domain in AI is to represent real-world phenomena or processes in a format suitable for computer utilization, primarily with the aim of facilitating automation (Surden 2011), and problem construction is the starting point for the execution of tasks by AI (Wirth and Hipp 2000). The process of problem construction involves semantic transformation, variable selection, and other procedures to translate abstract requirements into measurable, programmable, and observable quantitative features that can be processed by computers. For instance, in attempting to use AI to assist in filtering “crime suspects”, the target requirements may be divided into measurable features such as gender, age, occupation type, education level, criminal record, etc. Subsequently, a systematic and extensive screening process can be conducted based on these features.

For the platform of algorithm design, the primary goals of AI applications are often linked to efficiency and benefits. Even though there is an increasing emphasis on the direction of “fairness”, the concept of fairness may remain elusive, preventing its incorporation into the design framework of AI. Particularly when confronted with pre-existing discrimination in society, such as discrimination based on race, gender, education level, etc., which may manifest due to human negativity (e.g.stereotypes, biases) and unintentional prejudices (e.g., organizational practices or inherent stereotypes), it may enter algorithmic learning through the screening of critical computer variables. In comparison to traditional forms of discrimination, algorithmic discrimination resulting from problem construction exhibits more abstract, less intuitive, subtle, and challenging-to-detect characteristics.

3.2.2 Algorithmic Discrimination Risk in the Dimension of Data Processing

Data is the core of algorithms, and the representativeness, objectivity, and direction of data feature selection may profoundly impact the efficiency, accuracy, and fairness of algorithms. If algorithms are trained on inaccurate data (Pauline 2016), biased data (Barocas and Selbst 2016), or unrepresentative input data (Suresh and John 2019), they may generate discriminatory or biased outcomes. Therefore, there are currently three main reasons in the field of data processing that may lead to algorithmic discrimination:

1. Numeric Discrimination Caused by Insufficient Data Representativeness:
   In the current data era, data is primarily generated through passive, active, and automatic means, namely, relying on human-generated passive records, user-dependent autonomous creation, and perception system-dependent automatic generation (Meng and Ci 2013). However, due to the inherent imbalance in social group structures, some groups may lack internet terminal tools, have difficulty mastering popular internet technologies such as Weibo and WeChat, or be restricted by infrastructure limitations in the Internet of Things. As a result, the data related to them is inevitably influenced by these subjective and objective factors, leading to omission and neglect. The data of these specific groups becomes a hidden or even blind spot in the data analysis process. For example, as of the end of 2023, the population of individuals aged 60 and above in China has reached 290 million,[1] but the corresponding number of elderly internet users is only 140 million.[2] This implies that approximately 150 million elderly individuals have almost no digital footprint in the current internet, leading to a high likelihood of missing elderly data.
2. Numeric Discrimination Caused by Excessive Data Representativeness:
   Corresponding to insufficient data representativeness is the excessive processing of data for specific groups, resulting in disproportionate over-representation. Taking the predictive policing application, which has already reached a certain scale of application, as an example, it utilizes statistical methods to correlate existing crime situations with characteristics of criminals, thus delineating crime levels among different groups and serving relevant criminal judicial adjudication (Harcourt 2006). In the judicial practice of criminal litigation, the average arrest rate for the floating population is usually higher than that for local registered residents, while the bail rate is lower than that for local registered residents (Zhang 2014). The main reasons for this situation are the inherent mobility of the floating population, differences in regional administrative supervision methods, and the ambiguity of arrest usage rules. However, if the variables of the floating population are directly taken as representative data, it is highly likely to form a judgment of “higher crime rate among the floating population.” Therefore, excessive data representativeness may lead to excessively dense phenomena in

[1] See National Bureau of Statistics: “Statistical Bulletin of the People’s Republic of China on National Economic and Social Development for the Year 2023”, National Bureau of Statistics official website, https://www.stats.gov.cn/sj/zxfb/202402/t20240228_1947915.html (accessed on March 12, 2024).

[2] See 51st Statistical Report on Internet Development in China, China Internet Network Information Center, https://cnnic.cn/n4/2023/0302/c199-10755.html (accessed on March 23, 2024).

the data for certain groups, thereby triggering data hotspots. When data itself is excessively concentrated, resulting in imbalanced representativeness, the machine itself lacks a recognition mechanism. Thus, based on a foundation of “bias”, it will ultimately reproduce or even exacerbate discriminatory behavior that already exists in real life.

3. Numeric Discrimination Caused by Improper Data Feature Selection:
   The normal operation of AI typically relies on programmers setting specific target requirements as appropriate features, thereby depicting the correlation between independent and dependent variables. Feature selection involves removing redundant or irrelevant features from a set of features (Wang, Linlin, and Yao 2005), finding the truly influential ones, and seeking the optimal solution among model efficiency, accuracy, and cost. As Barocas and Selbst claimed that an organisation must make choices about what attributes they observe and subsequently fold into their analyses (Barocas and Selbst 2016). However, when the number of candidate data features is insufficient or excessive, it may lead to problems such as the curse of dimensionality, overfitting, and noise invasion, thereby reducing model performance (Cui et al. 2018).
   Generally, when it comes to the selection of data features, there are two primary methods: automatic learning by algorithms and induction based on human experience (Molnar 2020). If the feature set includes legally protected identity information such as gender, race, and physical health status, it may lead AI to make judgments based on the accumulated feature values, triggering digital discrimination. Therefore, in the development process of judicial AI, it is necessary to handle protected collective identities in a special manner to ensure the safety and fairness of AI use. However, although this design approach appears to avoid the potential risks of algorithmic discrimination on the surface, it is challenging in substance to eliminate the correlation between collective identity and candidate features. As a result, discriminatory outcomes may arise in AI applications that seem “fair” and “just”, This is especially evident in the field of AI-assisted judicial adjudication, where the attributes of the parties involved are more pronounced compared to other everyday scenarios. If the feature selection in judicial AI is not well-handled, it is highly likely to violate the principle of “equality before the law” in judicial procedures, leading to algorithmic discrimination against minority groups and even differences in race and gender.

3.2.3 Algorithmic Discrimination Risk in the Dimension of Algorithmic Logic

The underlying logic of algorithms is inductive, meaning that it derives corresponding “experience” from existing datasets and uses the learned “experience” to achieve the effect of judgment. The process can be described as follows: first, collect a sufficient amount of data, then filter, clean, and classify the data for training. The predefined problem needs to be divided into several related sub-problems, and programs and data are utilized to train reasonable models. This enables the processing of new data in the future. Therefore, when the data itself contains discriminatory factors or when discriminatory factors are set as labels, AI is highly likely to perpetuate and deepen discrimination through learning and reinforcement of this digital technology.

One of the most well-known cases in which AI assists judicial systems is the investigation into COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). COMPAS is an algorithm used by judges to assess the likelihood of a defendant reoffending (Angwin, Larson, and Mattu 2022). It employs a classification approach, extensively learning from existing crime facts, and assists police in predicting the likelihood of a particular subject committing a crime, thereby aiding in public safety. However, the non-profit organization Pro Publica, through an investigation, found that the algorithm was deemed favorable to white defendants: for example, among individuals re-arrested, the likelihood of white defendants being incorrectly classified as low risk was nearly twice that of black defendants (false negatives). Conversely, the likelihood of black defendants not re-arrested being incorrectly classified as high risk (false positives) was nearly twice that of white defendants. Therefore, the classification thinking of AI-assisted judicial sentencing algorithms, which involves collecting, extracting, classifying, and summarizing universally applicable rules from existing training sets to guide practice (Paul and Edward 1982), presents the potential for societal discrimination and exacerbates existing risks in society.

3.3 Injustice Within the Computational Environment

3.3.1 Algorithmic Hegemony: The Unfair Distribution of Data

The emergence of judicial AI as a tool for decision-making in legal contexts has introduced both opportunities and challenges, and its core is to enhance efficiency, consistency, and objectivity in legal proceedings by using data-driven algorithms to predict outcomes such as sentencing, parole decisions, or bail assessments (Susskind 2023). However, this reliance on algorithms raises significant concerns about algorithmic hegemony, a phenomenon in which the concentration of power over data and the control of algorithmic processes leads to systemic bias and injustice. Algorithmic hegemony in judicial AI is not merely a technical issue; it is fundamentally a social, political, and legal one, as it reflects the uneven distribution of power and resources across different social groups. When powerful entities, such as governments, private tech companies, or dominant legal institutions, control both the data used to train these systems and the algorithms that process it, the result can be a reinforcement of existing biases and inequities. In this context, algorithmic discrimination emerges not only as a technical flaw but as a reflection of the unequal distribution of data ownership, control, and influence in shaping legal outcomes. Understanding algorithmic discrimination within the broader framework of algorithmic hegemony is crucial for addressing its ethical and legal implications.

Firstly, algorithmic hegemony manifests in the centralization of data ownership. Data ownership pertains to the legal rights and control over the data used to train judicial AI systems. The ownership of legal datasets by private entities or state bodies determines whose interests are represented in algorithmic decisions. When data ownership is concentrated in the hands of a few, particularly powerful entities – such as large technology firms or government agencies – those entities wield disproportionate influence over the design and operation of AI systems. Data ownership touches upon key issues of property rights, sovereignty, and individual autonomy. As Shoshana Zuboff argues in *The Age of Surveillance Capitalism*, when corporations control vast amounts of data, they not only dominate market resources but also shape societal norms and political power (Zuboff 2023). In the context of judicial AI, this control over legal data means that decisions about who gets to access justice – and in what form – are made by private actors whose interests may not align with those of the public. For example, historical case data used in AI training may overrepresent the experiences of certain demographic groups while underrepresenting others, leading to injustice in legal outcomes. Data monopolies, therefore, concentrate power and shape the legal system in ways that benefit a few while marginalizing many.

Secondly, algorithmic supremacy is embodied in data control, the ability to shape judicial AI outcomes. Data control involves not only the ownership of data but also the power to shape its collection, curation, and use within AI systems. Control over data dictates how data is selected, which data points are prioritized, and how data is interpreted. In the judicial context, entities with control over legal data – whether government institutions or private corporations – can influence the development and functionality of AI algorithms in ways that may perpetuate discrimination. The bias in data control can manifest in several ways, from underrepresentation of certain groups to the reinforcement of existing prejudices. For example, if judicial AI systems are trained on data that reflects historical biases – such as racial disparities in arrests and sentencing – the algorithms will likely replicate these patterns, resulting in discriminatory judgments. Ruha Benjamin's Race After Technology (Benjamin 2019) highlights how the control of data not only shapes technology but also reinforces the marginalization of racial minorities. AI systems are often deployed in legal settings with insufficient checks on data quality, and without an accurate, representative dataset, the system perpetuates the biases embedded in historical data. Data control in judicial AI also ties into concerns about algorithmic accountability. In his discussion of regulatory capture, Frank Pasquale in *The Black Box Society* argues that the entities controlling algorithmic systems have little incentive to ensure fairness or transparency (Pasquale 2015). This lack of oversight means that individuals subjected to AI-driven legal decisions often lack the means to challenge biased or discriminatory outcomes. The concentration of data control thus becomes an issue of fairness in the judicial system, as it gives disproportionate power to those who curate and interpret data, shaping not only the algorithmic logic but also the legal conclusions derived from it.

Finally, algorithmic supremacy is embodied in data power, the ability to influence and shape legal structures. Michel Foucault's theory of power in *Discipline and Punish* suggests that systems of control often operate in subtle ways, shaping behavior without direct coercion (Foucault 1975). In the case of judicial AI, data power operates by influencing judicial outcomes and societal norms through the decisions made by AI algorithms. These algorithms often have outsized influence on public opinion, media narratives, and policy frameworks, shaping perceptions of who is deserving of justice and who is not. This power is further exacerbated by the openness of judicial AI systems to commercial exploitation, where decisions made by these systems may not only be biased but also influenced by market interests. Additionally, data power can reinforce existing societal inequalities. If judicial AI systems are primarily trained on data that reflects historical biases against certain racial or socio-economic groups, the algorithms will inevitably perpetuate these biases. David Garland in *The Culture of Control* (Garland 2002) discusses how surveillance technologies, including AI, become tools for maintaining social order by reinforcing inequalities. In the context of judicial AI, this creates a feedback loop where those already marginalized by the legal system are disproportionately targeted and penalized. The data power embedded in judicial AI is thus both structural and pervasive, extending far beyond the individual decisions of courts to shape societal norms and legal standards. The power to influence legal outcomes is not just about the algorithms themselves but about how data-driven systems become entrenched in the broader legal infrastructure, reinforcing power dynamics that benefit those who control the data.

3.3.2 Procedural Injustice: Lack of Transparency and Accountability in AI Systems

Algorithmic discrimination in judicial AI is deeply intertwined with the opacity of procedural injustice. When AI systems are used in judicial decision-making, such as sentencing or parole assessments, they often operate under a veil of secrecy, where the logic behind their decisions is hidden from both the public and the individuals affected (O’Neil 2016).This lack of transparency in algorithmic processes is not a mere technical shortcoming; it represents a procedural injustice that undermines the right to a fair trial and the rule of law. Discrimination by judicial AI may occur when biased data or flawed algorithms disproportionately affect certain groups, but the opaque nature of these systems makes it exceedingly difficult for those impacted to challenge or understand the basis for these biased outcomes. As a result, marginalized individuals may experience unfair legal decisions without the ability to contest them. The opacity of AI decision-making is not only a technical issue but a fundamental violation of procedural fairness, where individuals are denied both the ability to scrutinize the decision-making process and the right to seek redress for possible biases.

Firstly, the opacity of the algorithm decision affects the fairness of the process. One of the core issues with judicial AI systems is the lack of transparency regarding how these systems make decisions. Many judicial AI algorithms, such as those used to predict recidivism risk or determine bail conditions, operate as black-box systems, meaning that their decision-making processes are not accessible or understandable to the public, legal practitioners, or even the individuals directly affected by these decisions. This lack of transparency prevents stakeholders from understanding the rationale behind specific decisions, undermining the principle of due process and fair trial rights. AI creates a fundamental disconnect between the individuals affected by AI-driven decisions and the decision-making process itself. If a defendant does not understand why an AI system has assessed them as a higher or lower risk, they are unable to challenge or contest the decision, which directly impinges on their right to a fair trial. Furthermore, data-driven systems in judicial contexts often rely on historical data that may be flawed, incomplete, or biased. When the decision-making process is opaque, it becomes virtually impossible to discern whether a given decision is based on prejudicial data (e.g., racially biased arrest records) or statistical correlations that do not account for the complexity of human behavior and legal nuance. For example, the COMPAS system used in the U.S. has faced scrutiny for predicting recidivism risk based on data that may reflect systemic racial biases, but without transparency into how the model processes this data, individuals cannot fully understand or challenge the algorithm’s predictions. The absence of transparency in these processes means that discriminatory patterns are perpetuated without meaningful opportunities for correction or legal challenge (Angwin, Larson, and Mattu 2022).

Besides, the imperfect accountability mechanism affects the realization of procedural justice. Closely linked to the opacity of judicial AI is the absence of accountability in the design, deployment, and outcomes of AI systems. While human judges are accountable through mechanisms such as appeals and judicial oversight, AI systems often lack such safeguards, making it difficult for affected individuals to contest or challenge decisions that may have been influenced by algorithmic biases. In many jurisdictions, there is insufficient legal oversight regarding the deployment of AI tools in legal proceedings. The lack of regulatory frameworks for AI decision-making creates a "regulatory vacuum" where decision-makers are not held to the same standards of accountability as human judges. This absence of accountability allows for the unchecked use of potentially biased or unjust algorithms in high-stakes legal decisions, such as sentencing, parole, and even the determination of guilt or innocence. From a legal theory perspective, the rule of law demands that individuals have the ability to challenge decisions that affect their fundamental rights. The lack of accountability in judicial AI directly undermines this principle. Max Weber's concept of rational-legal authority suggests that legal systems should be based on predictable and transparent rules, applied impartially (Weber 2019). However, when AI systems are deployed without clear accountability structures, their decisions can appear arbitrary or inconsistent, especially when the systems operate without proper scrutiny. This undermines not only the predictability of legal outcomes but also the legitimacy of the judicial system itself.

4 Regulatory Approaches to Mitigating Algorithmic Discrimination in Judicial AI

4.1 Clarifying the Limits of Algorithmic use in Judicial Decision-Making

4.1.1 The Limits of Formal Rationality in Judicial AI Systems

Formal rationality is a core concept proposed by Max Weber, emphasizing the systematization, calculability, and procedural nature of social systems. As for legal system, formal rationality is a hallmark, characterized by the internal consistency, logical coherence, and universal applicability of legal norms (Weber 2019). By operating through pre-established rules, formal rationality ensures the objectivity and predictability of legal decisions while excluding the influence of ethical, political, or personal factors. In judicial AI, the realization of formal rationality is primarily achieved through algorithms and data processing, enabling the automated application of legal rules. First, the application of systematic rules plays a central role. Judicial AI applies legal rules in a structured and consistent manner, following predefined algorithms to ensure that the application of legal norms remains unaffected by personal or emotional factors. For instance, AI models trained on case law utilize codified rules to assess facts, identify relevant precedents, and propose outcome recommendations, all while adhering to predefined algorithms. This procedural consistency ensures that the application of legal norms is insulated from subjective influences.

Additionally, the implementation of predictability and calculability is another key mode through which judicial AI embodies formal rationality. By generating legal decisions through algorithms based on inputs such as legal texts and case data, judicial AI reduces uncertainty in legal processes by calculating outcomes in advance. This capability enables stakeholders to better anticipate potential results, aligning with Weber's vision of a calculable legal order. For example, Shanghai's "206 System" incorporates features such as intelligent trial assistance, evidence verification, and intelligent support functions to achieve the goal of utilizing AI systems to assist with fundamental judicial tasks. This system enhances efficiency in handling routine judicial work and reasoning in straightforward cases, while also contributing to the judicial objectives of "consistent rulings for similar cases" and ensuring "fairness and objectivity." The intelligent voice system aids judges in recording court proceedings and provides real-time transcription of trial processes to ensure their accuracy and fairness. The intelligent evidence verification function analyzes evidence to automatically identify and exclude false testimony and illegal evidence, thereby safeguarding the validity and objectivity of trial outcomes. Meanwhile, the intelligent support function assists judges in reasoning and decision-making by analyzing case information and relevant legal provisions, improving the rationality of judicial rulings.

While judicial AI demonstrates numerous advantages in terms of formal rationality, its limitations cannot be overlooked. Excessive formal rationality may lead to a legal system that is detached from ethical or contextual considerations. This concern manifests in the field of judicial AI as the rigid adherence to formal rules by machine learning systems, often at the expense of substantive justice in individual cases. In particular, the training data extracted from historical legal records may contain biases or discriminatory patterns. These biases can be learned and perpetuated by AI systems, leading to unfair treatment of specific groups and resulting in formalized, patterned algorithmic discrimination (Barocas and Selbst 2016).

Moreover, the "algorithmic black box" and "automated decision-making" of AI models increasingly drive the emergence of their "dehumanized" characteristics. Just as Marx, during the rapid development of capitalism, profoundly reflected on the risks of human "objectification" and the possibility of humans being replaced by various mechanical tools, we can similarly reflect on the risks of human "digitalization" in the current digital age. Digital algorithms have replaced biological algorithms in the learning, cognition, judgment, and decision-making processes of data-driven practical tasks.

With the amplification of formal rationality through algorithms, human rationality is gradually being supplanted by the logic of “efficiency” and “reasonableness,” as algorithms increasingly handle specific tasks and achieve specific goals more effectively. In this context, the human factor is relegated to the “object” position, subjugated within the system. The model of democratic development based on intersubjectivity, as envisioned in Jürgen Habermas’ concept of “communicative rationality”, becomes particularly challenging in the application scenarios of AI. The ability to build a “communicative community” through dialogue, understanding, and consensus is at risk of being eroded. Habermas criticized the limitations of formal rationality in legitimizing legal systems, particularly its separation from democratic deliberation and moral substance (Habermas 1996).This perspective also highlights the need to recognize the limitations of judicial AI and to carefully manage its relationship with humans, as the blurring of boundaries between humans and AI rooted in the manner in which humans situate AI (Cheng and Liu 2023). In the process of algorithm design, reliance should not be placed solely on abstract legal rules; instead, algorithms should be capable of adapting to societal changes and practical needs. Algorithms must possess a certain degree of flexibility and adaptability, allowing them to adjust in response to evolving social contexts and new moral standards. Through continuous learning and updating, algorithms should reflect the latest social developments and moral consensus, thereby reducing discriminatory outcomes.

4.1.2 The Limits of Substantive Rationality in Judicial AI Systems

The concept of Substantive Rationality rooted from Immanuel Kant’s moral philosophy. It was initially understood as the alignment of rational actions with universal moral imperatives. For Kant, rationality was not merely about consistency in thought or action but also about adhering to ethical principles that could be universally applied. This form of Substantive Rationality emphasizes moral reasoning as a key component of rational decision-making, where the justification of actions must extend beyond mere logical coherence to encompass the ethical worth of those actions (Kant 1998). Max Weber, building on the ethical framework of Kant, adapted Substantive Rationality to his analysis of legal and bureaucratic systems. Weber distinguished between “formal” and “substantive” rationality, arguing that while formal rationality is concerned with the logical consistency of rules, substantive rationality involves the application of those rules in ways that reflect deeper moral and social values (Weber 2019). The evolution of Substantive Rationality continued with Ronald Dworkin’s theory of law, where he argued that legal decision-making should not be solely driven by rules or procedures, but by moral principles that reflect society’s commitments to justice. Dworkin’s “law as integrity” asserts that legal decisions should interpret rules in a way that makes them the best possible reflection of community values (Dworkin 1986).

When applied to the issue of algorithmic discrimination in judicial AI, the theoretical evolution of Substantive Rationality underscores the need for a legal system that considers both procedural fairness and substantive justice in the face of automated decision-making. Besides, there is a close relationship between substantive rationality and the activism of the judiciary. Substantive rationality primarily emphasizes the substantive justice of judicial decisions, while judicial activism emphasizes the subjective initiative of judges in adjudication, i.e., the judgment and decision-making freedom exhibited by judges when facing specific cases. Judges need to consider various factors such as facts, law, morality, fairness, etc., and there are complex relationships and balances among these factors. Faced with complex cases, judges need to evaluate and interpret based on different value systems, and this evaluation and interpretation often rely on the judges' professional knowledge and experience and cannot simply depend on quantitative data and logical models. Therefore, in the current digital age, big data, and AI technologies still have certain gaps in achieving substantive rationality in judicial practice (Eubanks 2018).

Certainly, legal positivism provides strong support for AI algorithms simulating legal reasoning, and it can make a defense to some extent, as mathematical sciences inherently have advantages in formalization. If legal studies can accept these methods as their models, success similar to mathematics can be achieved. However, such theoretical assumptions cannot achieve such precision in real life. In concrete practice, judges need to repeatedly assess the relevant facts and applicable legal rules in the face of complex cases. Moreover, for written laws, they can make different interpretations based on different situations. This process of value measurement is a crucial basis for fair judgments and substantive rationality. However, the current development of judicial AI finds it challenging to measure a fixed set of criteria across different value systems, making it difficult to determine how to make reasonable judgments on specific cases.

In recent years, judicial authorities have gradually recognized the drawbacks of such "mechanical" sentencing. However, how to address the issue of "mechanical" sentencing, especially with the intervention of AI in the judicial field in the digital age, remains a topic that needs constant exploration. Scholars such as Ji Weidong also point out that in criminal proceedings, which involve "life, freedom, national goals, and social justice", the possibility of human sentiment and reformation should be retained. However, these potential factors are challenging to fit using technology (Ji 2007).

As Professor Yu Xingzhong pointed out, "AI is the crystallization of human intelligence, capable of reflecting human thinking to some extent. However, humans have not only intelligence but also emotions and spirituality, and these aspects complement and balance each other" (Yu 2016). AI may simulate human intelligence in specific judicial practices, making predictions based on data. Still, given the current intensity of machine learning, it is challenging to simulate the emotional and spiritual aspects of humans. Therefore, based on a careful examination of AI in terms of formal rationality and substantive rationality, it is necessary to define its limitations and clarify its scope of application. In the process of building AI to assist judicial rulings, it is crucial to emphasize the logical combination of formal rationality and substantive rationality, integrate formal and substantive reasons, strengthen supervised machine learning methods, and provide a mechanism for judges to give timely feedback and corrections. This approach will better train AI, especially in aspects involving argumentative reasoning and value judgment in judicial decision-making activities. Additionally, safeguards should be put in place for judges' decision-making authority, focusing on the human capacity for sentiment as a moral and ethical individual, to truly and reasonably apply judicial AI.

4.1.3 Classifying Judicial AI Applications and Determining Appropriate Boundaries

Cardozo pointed out that if "the points involved in a case are the same, the parties expect the judge to make the same decision" (Cardozo 1928). For two cases with the same or similar circumstances, judges should equally apply the law and make the same or similar judgments, known as "same case, same judgment". This demonstrates equality in the application of the law and is at the core of justice. Therefore, based on the characteristics of the algorithm itself, the author categorizes the application scenarios of judicial AI into the following three types. It is hoped that by defining the boundaries of algorithmic use, the expected effect of "same case, same judgment" can be better achieved, and the rational defects brought about by judicial AI can be minimized as much as possible.

1. Automated Decision-Making for Simple Factual Cases.
   Judicial AI, especially intelligent adjudication systems, is a significant trend in the modern legal field, particularly in cases with straightforward facts and undisputed circumstances. In such cases, the necessity for judges' discretionary judgment is relatively low. Therefore, intelligent adjudication systems can provide judges with accurate judgment predictions, enabling fully automated decision-making, thereby enhancing judgment efficiency and accuracy. For instance, in cases processed through simplified procedures, if there is sufficient evidence to prove the defendant's criminal acts, and the defendant has already signed a confession and acceptance of punishment, judicial AI can be utilized for assistance by pre-setting sentencing algorithms, avoiding unnecessary human

interference and errors, thus improving the efficiency and consistency of judgments.

Moreover, in situations where less experienced judges confront cases lacking specific standards, intelligent adjudication systems can offer them insights into how similar cases were handled and their outcomes, thereby standardizing the exercise of judges' discretionary powers. By dealing with numerous simple and typified cases, intelligent adjudication systems can effectively reduce bias and injustice arising from preconceptions, promoting fairness in judgments. Legal or judicial decisions should be predictable, meaning they should be based on previously published general legal rules (Lei 2015). In comparison to traditional judicial decision-making models, judicial AI needs to enhance its interpretability and predictability to maintain the stability and reliability of the judicial system, as demonstrated effectively in simple factual cases. However, it is crucial to note that intelligent adjudication systems cannot entirely replace the role of judges. In complex cases with disputed facts, judges still need to rely on their professional knowledge and experience for comprehensive analysis and decision-making, where AI's judgment and decision-making capabilities cannot fully substitute.

In conclusion, intelligent adjudication systems can adopt fully automated decision-making processes in cases with straightforward facts and undisputed circumstances, thereby improving the quality and efficiency of judgments and propelling the operation of judicial trial processes. Nevertheless, with the rapid development of the digital age, the increasing application of intelligent adjudication systems in the legal field also requires continuous reinforcement of supervision and application of AI technology to ensure its accuracy and fairness in judicial practice.

2. Semi-Automated Decision-Making in Complex Cases

   Difficult cases typically refer to those requiring judges to exercise a certain degree of subjective initiative during the case-handling process. These cases are generally fewer in number and involve challenging value judgments. At the current stage, judicial AI lacks the ability to derive experiences from extensive learning, especially when dealing with conflicting legal rules or scenarios involving challenging value judgments, making it more difficult to achieve balance (Hart 2012). Therefore, in such cases, AI can only provide judges with certain references through its own case retrieval and recommendation functions as a preliminary step to assist judges in making judgments. In this sense, we categorize the judicial AI's role in this type of scenario as semi-automated decision-making. However, the number of complex and difficult cases is currently relatively small, and the resources for learning databases are limited. Therefore, the accuracy of case recommendations still needs to be examined. In the existing judicial context, except for some newly appointed judges, the demand from most judges for system-recommended similar cases is not substantial. Consequently, compared to other scenarios, the usage of judicial AI is more restricted in difficult cases, emphasizing the need for judges to rely on their own experience and knowledge to make appropriate judgments.

3. Assisted Decision-Making in Risk Prediction Scenarios
   For cases with simple facts and no disputes over factual determinations, different judges generally have no divergence in their judgments. In such cases, judges have minimal to nonexistent discretionary power. The extensive deployment of risk prediction AI is not fully realized at this stage. Therefore, addressing the potential types of judicial AI that may emerge locally in the future holds both innovative and meaningful implications. Furthermore, due to the triple dilemma of biased data inputs, reinforcement mechanisms in algorithm design, and the algorithmic black box issue during program execution in predictive AI, there is an increased likelihood of exacerbating discriminatory algorithmic practices in the field of judicial adjudication. As Li Xunhu pointed out, "The legal problem with the application of AI in the criminal judicial adjudication field lies in: intelligent case handling systems will exacerbate existing biases", and in specific applications of crime risk prediction, it can be categorized as "the use of intelligent risk assessment tools will result in discrimination against specific groups (Li 2021)". Therefore, it is necessary to contemplate foreseeable risks of algorithmic discrimination in local judicial AI.

   In current AI applications for crime risk assessment, a typical example is the COMPAS system in the U.S. judicial practice. The system has been criticized for exhibiting algorithmic discrimination as it distinguishes between racial attributes in predicting recidivism rates (Feller et al. 2016). Specifically, under similar conditions, the likelihood of a Black person reoffending is twice that of a White person, but such discriminatory predictions cannot be substantiated in actual judicial practice.

   Hence, for risk prediction scenarios in judicial AI, its utility in judicial affairs should be auxiliary and must adhere to the requirement of "interpretability". The characteristics of auxiliary decision-making necessitate users of judicial AI to consider the data and conclusions provided by the AI within limits and rely more on their own judicial practice experience to make judgments. The interpretability requirement specifies the operational mechanism of the algorithm, ensuring human understanding of its decision rules. It also explains why a particular judicial decision is made in specific professional contexts, clarifying the reasonableness of its output results.

4.2 Diversifying Regulatory Approaches to Judicial AI Governance

4.2.1 Strengthening Legal Frameworks for the Regulation of Judicial AI

The regulation of AI in global judicial systems varies significantly. The United States, with its legislative characteristics of openness and inclusivity, adopts a more nuanced and targeted approach to regulate AI applications in different scenarios, focusing on small-scale, precise regulations. This strategy avoids comprehensive national legislation, aiming to prevent the stifling of technological development. In contrast, the European Union employs a unified, broad legislative model, which exerts widespread and profound influence in the international arena. Through legislative measures, the EU seeks to establish a significant voice in the global discourse on AI.

In the United States, the regulation of AI can be analyzed at both the federal and state levels. At the federal level, despite numerous AI-related proposals, no comprehensive AI legislation has been passed to date, and there is a lack of specific, comprehensive laws addressing algorithmic discrimination. AI regulation primarily relies on executive orders issued by the President, focusing on areas such as national security, military, and foreign relations. For example, Executive Order 13,859, *Maintaining American Leadership in Artificial Intelligence*[3] (issued by the Trump administration), and Executive Order 13,960, *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*[4] (also under Trump), mainly regulate the use of AI by U.S. regulatory agencies. Similarly, Executive Order 14,110, *On the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*[5] (issued by the Biden administration), requires federal regulatory agencies to establish standards and norms for AI use in sectors such as criminal justice, education, healthcare, housing, and labor.

In comparison, state-level legislation is generally easier to pass and tends to focus on areas such as consumer protection, anti-discrimination, and civil rights protection, with limited regulatory obligations. For example, Colorado's regulation of

[3] See The White House, Maintaining American Leadership in Artificial Intelligence, https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence (accessed Jan 21, 2025).

[4] See the White House, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government (accessed Jan 21, 2025).

[5] See The White House, On the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence (accessed Jan 21, 2025).

high-risk AI systems emphasizes consumer protection and resulted in the passage of the *Colorado Artificial Intelligence Act*, which is considered the first state-level legislation in the U.S. to regulate high-risk AI. Similarly, the New York City Council passed Local Law No. 144:Automated Employment Decision Tools (AEDT),[6] which prohibits employers and employment agencies from using AI and algorithm-based technologies for recruitment, hiring, or promotion without conducting a bias audit.

Overall, the United States has maintained a high level of inclusivity in the fields of technology and innovation. The market-driven and innovation-oriented legislative approach has facilitated business development and market competition, as evidenced by the positive role of law in the development of sectors such as the internet and information technology. In terms of risk regulation, the U.S. has focused on regulating the use of AI by government agencies to ensure public trust and protect citizens' rights, while also applying existing laws to impose limited regulation on businesses, thereby avoiding excessive market intervention. However, the lack of a unified federal AI legislation may result in regulatory fragmentation, making it difficult to establish a comprehensive and effective regulatory system. State-level legislation has, to some extent, addressed the issue of algorithmic discrimination, providing certain regulations for AI applications in specific sectors. However, the significant variations in state-level laws, along with the lack of coordination, may lead to businesses facing differing compliance requirements in different states. This also complicates the creation of a comprehensive and systematic framework for addressing algorithmic discrimination, resulting in regulatory gaps and inconsistencies.

The European Union has adopted a distinctly different legislative approach to AI compared to the United States. It pursues a unified regulatory approach through the *Artificial Intelligence Act*,[7] which standardizes and classifies the regulation of AI risks. This legislation covers both product-based and decision-support AI systems and assigns specific responsibilities to different stakeholders based on risk levels, aiming to reduce algorithmic discrimination risks to some extent through risk classification and tiered governance. In terms of risk categorization, the EU's *Artificial Intelligence Act* introduces a unified framework, classifying risks into four categories: prohibited, high-risk, limited-risk, and minimal-risk. The legislation places particular emphasis on regulating high-risk AI systems. For instance, high-risk

[6] See The New York City Government, Automated Employment Decision Tools (AEDT), https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page, (accessed Jan 21, 2025).

[7] REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, (accessed Jan 21, 2025).

AI systems are required to provide high-quality datasets to mitigate discriminatory outcomes. With the explosive growth of generative AI and large models in recent years, the Act has proposed a revision that introduces the concept of "general-purpose AI models," which are subject to special regulations. Providers of such AI models are responsible for submitting technical documentation, disclosing information and documents to downstream AI system providers, providing summaries of content used in model training, and cooperating with regulatory authorities.

The European Union's unified regulation of AI helps establish relatively consistent regulatory standards across the EU, facilitating coordination among member states. In particular, its focus on protecting vulnerable groups aims to uphold social fairness through corrective justice, thereby reducing the risks associated with algorithmic discrimination. However, the risk regulation approach overly emphasizes formal uniformity, resulting in somewhat arbitrary risk classifications across different industries and sectors. For example, in addressing emerging technologies like generative AI, there is a dilemma in categorization, making it difficult to keep pace with the rapid development and diversity of AI technologies. Additionally, given the EU's relatively weaker industrial and technological strength in the digital sector, its stringent legislation may pose significant obstacles to the development and innovation of digital technologies, hindering scientific and technological progress.

To effectively address the complex and increasingly severe issue of algorithmic discrimination in AI, particularly in the judicial sector, it is necessary to develop a comprehensive and detailed legislative strategy from multiple dimensions. First, drawing on the flexibility and targeted nature of U.S. legislation, a regulatory path with specific and focused approaches should be created to address key issues in the judicial AI sector. For example, laws and regulations should be formulated to address risks such as data leakage in judicial case management systems or bias in evidence analysis algorithms that may affect the evaluation of evidence for specific groups. This approach not only enables effective risk prevention but also provides the necessary space for the development and exploration of AI in the judicial field. In the current complex and rapidly evolving technological environment, this is the optimal choice, and likely will remain so for the foreseeable future.

In this process, it is essential to carefully manage the many dialectical relationships involved in the development of judicial AI. First, domestic and international coordination must be handled well. Domestically, a regulatory framework for AI algorithms should be developed that aligns with the national judicial system and legal culture. Internationally, it is crucial to actively participate in the formulation of international standards, strengthen exchanges and cooperation with other countries on judicial AI algorithm regulation, and enhance the nation's influence in this area. Second, the coordination between hard power and soft power must be balanced. In terms of hard power, there should be increased investment in the research and development of judicial AI algorithms to improve technological capabilities. Regarding soft power, attention should be paid to the integration of legal culture and the rule of law principles into algorithmic regulation, cultivating a positive international image. Third, the dialectical relationship between development and security must be carefully managed. Innovation in judicial AI technologies should be encouraged to enhance judicial efficiency and fairness, while ensuring that security concerns – particularly the prevention of issues like algorithmic discrimination – are prioritized to protect the integrity of the judicial system. Finally, the relationship between each country's unique national conditions and its international influence must be addressed. The regulatory strategy for algorithmic discrimination should be tailored to the country's judicial practices, while also sharing successful experiences globally to enhance the nation's influence in international judicial AI regulation and promote fairness and justice in global judicial systems.

4.2.2 Enhancing Ethical Oversight Through AI Governance Committees

The current market landscape is characterized by the monopolization of AI-related technologies by specific suppliers and enterprises. The short-term difficulty in overcoming digital technological barriers makes the government's use of AI applications highly dependent on suppliers. The market for AI technologies is increasingly concentrated, with a few dominant companies controlling the development and deployment of critical AI tools (Brynjolfsson and McAfee 2014). This dependence poses a significant challenge to the exclusive authority of traditional legal adjudication. Based on this phenomenon, if judges excessively rely on AI-assisted adjudication systems to make judgments, it will result in a dual structure of adjudication, where both AI and judges share authority. When a judge uses an AI-assisted adjudication system to handle cases and make judgments, the decision-making process is no longer solely based on the judge's interpretation of the law; instead, it is substantially generated collaboratively by programmers, data processors, and technology suppliers. Therefore, apart from the need for precise regulations targeting external enterprises to ensure the legality and rationality of the market environment and application scenarios for judicial AI, there is also a requirement to establish a systematic ethical mechanism within enterprises. This internal ethical framework aims to ensure compliance with relevant ethical requirements at the inception of AI design, thereby better guaranteeing the secure application of AI.

Within enterprises, the author suggests the establishment of a high-level and influential AI Ethics Committee. The committee should consist of members with interdisciplinary, cross-domain, and diverse backgrounds, forming an expert team that includes ethicists, legal experts, and technical specialists closely related to specific topics. The functions of the AI Ethics Committee in enterprises can be mainly divided into three aspects:

1. Development of Ethical Policies: The committee should formulate ethical regulations and policies to facilitate the implementation of rules. As AI continues to evolve, it is crucial that explicit ethical regulations and institutional policies are developed to ensure its responsible use (Jobin et al. 2019). Before the standardization and orderly development of AI ethical norms, it is necessary to enact explicit institutional policies that help research and development and user entities clearly define the boundaries and potential harms of AI use. Taking IBM as an example, they have a mature AI Ethics Board that has established guidelines and specific policies for the company's AI use, ensuring that all projects operate within existing tracks and systems, enhancing the awareness of values and responsibilities among personnel involved in AI applications. Besides, it's better for the committee to develop dynamic policy-updating mechanisms to adapt to the rapid evolution of AI technology. For instance, the European Union's AI Act, which proposes a risk-based approach to AI regulation, provides a valuable model. This approach allows for the continuous assessment and adjustment of ethical policies, ensuring they remain relevant and effective as technology advances. Therefore, the committee could establish a regular review process that includes stakeholder consultations, expert evaluations, and feedback from end-users, thereby creating a responsive and adaptive ethical framework.
2. Implementation of Ethical Assessment and Review Mechanisms: Based on established ethical regulations and policies, the committee should conduct targeted ethical assessments and reviews of various aspects, including technological development, product application, market promotion, and feedback evaluation. This involves comprehensive evaluation of AI from the frontend, middle-end, and backend, addressing ethical risks in algorithmic processes, reducing the likelihood of algorithmic discrimination, and mitigating its impact. To be more specific, ethical scrutiny at the R&D stage must focus on the fairness and reliability of data and algorithms used in judicial AI systems. Judicial AI, including predictive tools for bail, parole, sentencing, and risk assessment, is heavily dependent on historical judicial datasets. These datasets often reflect existing biases in judicial decisions, including disparities in sentencing based on race, gender, or socioeconomic status. If not carefully reviewed, these biases may be perpetuated or even exacerbated by AI systems (Zarsky 2016). Ethical assessment during the deployment stage must focus on the practical implications of judicial AI tools in courts and legal proceedings. Judicial AI systems, such as automated risk assessments or document analysis tools, are being increasingly used to assist judges, lawyers, and clerks. While these systems can improve efficiency, they may

introduce risks of undue reliance, lack of explainability, and reduced human oversight in critical judicial decisions. As judicial AI systems are integrated into courts and promoted for wide-scale use, ethical assessments must address issues of transparency, accountability, and public trust. Judicial systems cannot afford overstated claims regarding the “neutrality” or “accuracy” of AI tools, as these claims may mislead stakeholders and obscure ethical risks.

3. Enhancement of Ethical Education and Value Construction: The responsibilities of the Ethics Committee include not just externally regulating AI research and usage projects, but also providing internal normative guidance (Jobin et al. 2019). It should strengthen the establishment of clear and firm values and ethical awareness throughout various stages of AI development. For instance, Google has provided its employees with a technology ethics training course. Through classroom teaching and practical exercises, employees are trained in moral awareness and ethical literacy, reinforcing their risk awareness. This empowers them to better evaluate potential pitfalls in AI and manage coordination mechanisms after identifying such risks.

4.2.3 Promoting Third-Party Auditing and Algorithmic Fairness Testing in Judicial AI

The fairness assessment of algorithms is a crucial consideration for the credibility of AI, with the primary goal of ensuring that algorithms, in their design and usage, do not discriminate against individuals or groups. Thus, facilitating fairness testing in algorithms is a significant measure to promote the secure use of algorithms and further advance the implementation of AI. Based on the machine learning algorithm lifecycle, mechanisms to eliminate algorithmic bias can be categorized according to the AI lifecycle, mainly involving preprocessing, in-processing, and post-processing. Firstly, when able to influence the data generation mechanism itself, a preprocessing mechanism can be employed to filter, clean, and synthesize data. Secondly, when it is possible to explicitly design the algorithmic process and control its operation, adjustments can be made in a manner that aligns with ethical standards of fairness and justice, such as adding constraints to machine learning models to eliminate potential bias. Lastly, when understanding the machine’s output mechanism and making reasonable adjustments, post-processing can be utilized to reinforce fairness (Liu 2021).

In the application scenarios of judicial AI, facial recognition stands as a common and critical component. Therefore, taking this typical scenario as an example, the author provides corresponding strategies for fairness testing and auditing. Currently, facial recognition technology is widely used in verifying individual identities, tracking fugitives, real-time video surveillance, and identifying victims.

How can a fair testing mechanism be established in the application scenarios of judicial AI to minimize the risk of digital discrimination? The author suggests addressing this from two aspects: data and algorithms.

1. Data Aspect: Since widely used facial recognition datasets, such as VGGFace 2 and MS1M, lack balance in attributes such as ethnicity and gender during their creation, models trained on these datasets inevitably introduce biases. To address this, targeted collection of more balanced datasets, such as the BUPT-Balancedface training set, can be employed to mitigate this issue.[8]
2. Algorithm Aspect: Even when trained on balanced datasets, algorithms can still produce biased outcomes, often due to complex interactions between features that might not be fully represented or accounted for in the training process. Therefore, addressing these issues directly in the algorithm is crucial (Barocas and Moritz 2023). For instance, Tencent Youtu Lab's "Consistent Instance False Positive Improves Fairness in Face Recognition" approach extends fairness concerns from the group level to the individual level. It achieves fairness by increasing the consistency of algorithmic False Positive Rates (FPR) at the individual level, thereby improving fairness at the group level.

Additionally, fairness testing can be conducted on algorithms after their design is completed. Fairness testing tools can be used to evaluate machine learning models after their design phase, enabling developers to assess whether the model produces biased or unfair outcomes in real-world applications (Sweeney 2013). For instance, IBM, a U.S.-based company, has introduced the IBM Watson Open Scale fairness testing tool to track and detect the output results of machine learning. This algorithm helps identify fair models even after the design phase, facilitating interpretable processing and compliant application. Microsoft has also developed the Fair learn toolkit to assist AI developers in autonomously evaluating whether their systems meet fairness requirements, making them closer to fairness before specific deployments. Therefore, given the vast application market for judicial AI, efforts can be made in two aspects. First, promote the universality and systematicity of algorithmic fairness testing. Various types of AI applications, especially those involved in designing public affairs and affecting individual rights in the judicial domain, should be included in the scope of algorithmic testing. Systematic testing across the stages of algorithm design – before, during, and after – can further ensure the security of AI.

[8] The BUPT-Balancedface training dataset is a racially balanced database, consisting of 7,000 individuals from each ethnic group. The BUPT-Globalface training dataset is constructed in proportion to the global population, maintaining the same ratios as the Earth's population for each ethnic group. This form of dataset construction is advantageous for addressing discriminatory issues in data sources.

Second, the algorithmic fairness testing mechanism should be more tailored to the characteristics of judicial scenarios. In addition to common ethical requirements such as “fairness” “justice” and “equality”, attention should be given to the features of different types of litigation, including civil, criminal, and administrative proceedings. This approach ensures more targeted fairness testing based on the specific characteristics of different legal contexts.

4.3 Promoting Justice and Equity in the Algorithmic Environment

4.3.1 Bridging the Digital Divide to Protect the Rights of Disadvantaged Groups

In the current era of ongoing digital civilization development, informatization, networking, and intelligence have brought us numerous dividends. However, data is gradually replacing traditional capital, becoming a tool to distinguish between social classes and hierarchies, and the emergence of the “digital divide” is continuously expanding. Due to the scarcity of data information itself and the subjective and objective gaps in individuals’ access to information, the benefits brought by the digital age cannot reach every member. Individuals lacking the ability to access data or use technology may be abandoned or entrapped by the digital realm, potentially falling into the category of “digital disadvantaged groups”. As data increasingly becomes a resource of paramount importance, it serves as the new capital that differentiates social strata, resulting in the expansion of digital inequality (van Dijk 2006).

In the field of judicial adjudication, a series of judicial transformations supported by AI technologies such as “online self-help filing” “online court” and “case recommendations” are gradually becoming popular. Modern technologies such as big data, cloud computing, AI, and blockchain are widely applied in litigation services, trial execution, judicial management, and other areas. In comparison to the digitization and intelligence seen in everyday life, the transformation in the field of judicial adjudication, due to its specialization and universality, further exacerbates the impact of the digital divide on vulnerable groups. Their rights and demands become even more challenging to deliver and safeguard.

Therefore, to create a just algorithmic environment, it is necessary to consider the inclusion of digitally disadvantaged groups in the digital society, safeguarding their rights, and mitigating the digital divide.

Firstly, adhering to the principle of equal protection is crucial. In the course of the current societal development, where the strong and weak coexist and the law of the jungle prevails, our pursuit of fairness and the protection of the vulnerable are important manifestations of civilization. Substantive justice requires that the legal system acknowledges the socio-economic disparities between individuals (Dworkin 2013), particularly in a digital society where some individuals may lack the capacity to engage fully with new technologies. Translated into the specific realm of citizens' information rights, this pursuit can be reflected in whether the law and society can, in the current development of the new technological revolution, safeguard the basic information rights of citizens in a fair manner. In the digital age, the challenge lies not only in ensuring equal access to technology but in addressing the profound asymmetry of power and resources between individuals, corporations, and the state, which necessitates a more nuanced approach to rights and obligations (Deibert 2020). This involves establishing a series of effective mechanisms for protecting rights, thereby maintaining social order in the era of digital civilization. Secondly, supplementing this is the assurance of preferential protection. In comparison to the primary principle of equality, preferential protection places greater emphasis on achieving substantive justice rather than the overt expression of formal justice. This theory has been extensively elaborated in the equality theories of scholars such as Dworkin and Rawls. In the digital age, individuals face an insurmountable gap in technological application capabilities compared to governments and corporations with significantly larger capacities. Therefore, in the formulation of rights and obligations, a corresponding balance should be struck (Gao 2019). Among different individuals, due to variations in the ability to access digital resources, it is necessary to initiate supportive policies for specific vulnerable groups. This might include ensuring that online and offline judicial operations run parallel and designating individuals for guidance and assistance.

4.3.2 Advancing Digital Equality as a Foundation for Algorithmic Justice

The root cause of AI discrimination risks lies in unconscious bias expression and structural inequality. Measures such as algorithm interpretation, algorithm auditing, and non-discriminatory compliance standards can reduce the probability of discrimination but cannot fundamentally shake structural inequality. Therefore, shaping a fundamental concept of digital justice in society and promoting digital equality actions are essential means to curb bias and structural inequality eroding AI (Li 2021).

Firstly, ensuring digital human rights and establishing the foundation of an equal rights movement are crucial. Digital human rights are emerging rights associated with the development of digital technology. The fundamental requirement of digital human rights is centered on people, with all data and technology based on human dignity and rights. Specifically, digital human rights should strengthen the protection of relevant rights such as information rights, privacy rights, data rights, and non-discrimination rights in daily life, reflecting the justice (Zhang 2019). For example, one of the most important facets of digital human rights is the protection of privacy. As digital systems collect vast quantities of personal data, individuals must have the ability to control their information. In many countries, privacy protections are rooted in laws such as the European Union's General Data Protection Regulation (GDPR), which provides individuals with the right to access, correct, and delete their personal data.

In addition, it is necessary to promote the diversity of AI application developers. If AI designers and model builders are exclusively dominated by a minority with obvious social characteristics, it may lead to the exclusion and discrimination of certain groups. Diversifying the development teams of applications can significantly incorporate the opinions of more common societal groups, facilitating more effective consultations and negotiations, ensuring the fairness and universality of algorithm development. Factors such as gender, race, economic status, educational background, and disciplinary background can be elements of diversity. Diverse personnel composition and organizational structure can guarantee that the logical design in the algorithm development process is not monopolized by one person. Different sources to a certain extent represent various groups in society, thus preventing algorithmic discrimination more effectively.

4.3.3 Enhancing Judicial Remedies and Strengthening Algorithmic Governance Mechanisms

Judicial relief, as the ultimate barrier in a rule-of-law society, plays a crucial foundational role in protecting citizens' information rights and pursuing digital justice. In a society governed by the rule of law, judicial relief serves as the last line of defense in ensuring the protection of citizens' rights, especially in an increasingly digitized world. As Richard Susskind argues in his seminal work *The End of Lawyers?* that the courts are the final arbiters in ensuring justice, and without effective judicial remedies, citizens' rights in the digital age are rendered vulnerable to exploitation (Susskind 2008). This underscores the necessity of judicial intervention in maintaining the balance between technological development and the safeguarding of fundamental rights. Therefore, in the process of digitizing justice, algorithmic governance-oriented judicial relief undertakes a significant mission. Due to the inherent integrative and systematic nature of algorithmic applications, it becomes especially critical to clearly define the rights and obligations and identify specific causal relationships in the process of specific rights protection and judicial relief.

Common judicial relief methods include the judicial review mechanism for algorithms. If litigation related to algorithms occurs, it is necessary for judicial authorities to initiate corresponding algorithmic audit procedures to facilitate the achievement of relevant judicial rulings. Once the algorithm has undergone an audit and issues are confirmed, it should be complemented by relevant ethical mechanisms to strengthen legal constraints and regulatory functions on algorithm platforms and designers. Specifically, clear provisions should be made for the responsible subjects, matters of responsibility, accountability standards, and modes of responsibility related to algorithms, helping to achieve more precise and clear governance of algorithms. Of course, while regulating algorithms, it is essential to protect their reasonable development process, preventing hindrances to the progress of algorithms and technology. Due to the inherent integrative and systematic nature of algorithmic applications, it becomes especially critical to clearly define the rights and obligations and identify specific causal relationships in the process of specific rights protection and judicial relief (Zuboff 2023).

In addition, enhanced protection should be applied to certain key areas closely related to fundamental rights, such as consumer antitrust and anti-unfair competition rights in algorithmic litigation. Special protection for digital vulnerable groups, with a particular focus on targeted safeguards for new-era digital human rights, is crucial. Based on these considerations, judicial relief is more likely to become the last powerful barrier, better upholding the principles of justice in the field of judicial AI applications and promoting the harmonious development of algorithmic governance. The judiciary need to act as a check on algorithmic governance, ensuring that algorithms do not operate outside the boundaries of established legal norms and that citizens' rights are upheld (Yeung 2018).

5 Conclusions

The exponential explosion of data and technology has ushered society into a new era of digital civilization, propelling the development of judicial modernization from the traditional digitization of case files towards various applications of judicial AI. Undoubtedly, the increasing prevalence of judicial AI applications has enhanced the efficiency of judicial operations, regulated judicial discretion, and continuously strived towards the uniformity and objectivity of legal application. However, the consequential risk of algorithmic discrimination looms large. Therefore, this analysis places a significant focus on the risk challenges of algorithmic discrimination in the field of judicial AI.

Firstly, the boundary of algorithmic application becomes blurred. The application of judicial AI disrupts the traditional authority in fact-finding and adjudication. The composition of legal provisions + judicial decisions has evolved into a synthesis of data engineers + software designers + judges. This transformation impacts the procedural justice and substantive justice inherent in judicial procedures.

Furthermore, the current use of judicial AI fails to effectively reconcile the balance between fairness and values. The automation of machine processing makes the analysis and judgment of cases mechanical and simplistic, easily neglecting the uniqueness of each case, thereby inducing the risk of algorithmic discrimination. Secondly, the pathways to algorithmic discrimination are diverse. The three major structural elements – problem formulation, data processing, and algorithmic logic – have the potential to introduce discrimination into AI algorithms within the judicial domain from various perspectives. This complex and covert nature of algorithmic discrimination necessitates more systematic and diverse regulatory measures to address the risks in judicial AI effectively.

Thirdly, the injustice in the algorithmic environment stems from the risks of data technology monopolies and the algorithmic tendencies of inherent societal discrimination. Therefore, for fundamental regulation of algorithmic discrimination in judicial AI, addressing the injustice at the foundational logic level is essential. This involves formulating ethical norms that align with the characteristics of the digital era.

Based on a multidimensional analysis of the challenges posed by algorithmic discrimination, the author conducted an in-depth investigation and analysis of judicial, technological, and ethical practices beyond the domain. A governance landscape for AI was delineated, and potential governance pathways were explored from three perspectives: legal regulation, technological responses, and ethical safeguards. Finally, the article, addressing the challenges presented earlier, constructed a systematic regulatory framework for addressing algorithmic discrimination in judicial AI, integrating knowledge from philosophy, law, and computer science. This framework unfolds across three major levels – algorithmic constraint specification, diversification of algorithmic regulations, and environmental justice of algorithms – and comprises nine detailed aspects.

First of all, algorithmic constraint specification encompasses the interpretation of “algorithmic rationality”, covering analyses of formal and substantive rationality constraints. Additionally, the author explicitly outlined a classification mechanism for specific judicial application scenarios, facilitating a more nuanced treatment of algorithmic application constraints.

Moreover, the diversification of algorithmic regulations includes the construction of legal regulations, ethical institutions, and the foundational aspects of technological fairness testing. It addresses discrimination in various aspects of judicial AI, providing point-to-point handling.

Lastly, the argument for the environmental justice of algorithms spans across bridging the digital divide, advocating for the protection of the rights of digitally disadvantaged groups, and actively promoting the digital equality movement, thereby solidifying the foundation of digital justice. Furthermore, the article emphasizes the role of judicial relief as the final line of defense to ensure timely governance following instances of algorithmic discrimination.

The application of AI is a major trend of the era, but this does not imply unrestricted, unregulated, or even harmful applications. For the judicial domain closely related to people's rights and social fairness and justice, the governance of AI becomes crucial, necessary, and urgent. This article analyzes the current issue of algorithmic discrimination in judicial AI as an entry point, combining perspectives from philosophy, law, and computer technology. It identifies present challenges and potential strategies, attempting to provide feasible insights for the governance of AI. The author acknowledges that resolving the issue of algorithmic discrimination in judicial AI is not an immediate task. Future efforts should involve specific practical validations, collection of opinions from various sectors of society, and a gradual approach to promoting the healthy development of AI.

Research ethics: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: This achievement is a phased outcome of the National Social Science Fund major bidding project: "Research on Accelerating the Modernization of Ecological Governance System and Governance Capacity" (No. 20 & ZD091).
Data availability: Not applicable.

References

Angwin, J., J. Larson, and S. Mattu. 2022. "Machine Bias. Ethics of Data and Analytics." *Auerbach Publications*: 254–64.

Arrow, Kenneth J. 1971. *The Theory of Discrimination*. Princeton: Princeton University Press.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104 (3): 671–732.

Barocas, Solon, Hardt Moritz, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge: MIT Press.

Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Medford: Polity Press.

Bentham, Jeremy. 1970. *An Introduction to the Principles of Morals and Legislation*. London: Athlone Press.

Binns, Reuben. 2018. "Fairness in Machine Learning: Lessons from Political Philosophy." *Conference on Fairness, Accountability and Transparency, PMLR* 81: 149–59.

Brynjolfsson, Erik, and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W. W. Norton & Company.

Cardozo, Benjamin N. 1928. *The Paradoxes of Legal Science*. New York: Columbia University Press.

Cath, C. 2018. "Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges." *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences* 376 (2133): 1–7.

Cheng, Le, and Xiuli Liu. 2023. "From Principles to Practices: The Intertextual Interaction between AI Ethical and Legal Discourses." *International Journal of Legal Discourse* 8 (1): 31–52.

Chesterman, Simon. 2020. "Artificial Intelligence and the Limits of Legal Personality." *International and Comparative Law Quarterly* 69 (4): 819–44.

Cui, Hongyan, Xu shuai, lifeng Zhang, Roy E. Welsch, and Berthold K. P. Horn. 2018. "The Key Techniques and Future Vision of Feature Selection in Machine Learning." *Journal of Beijing University of Posts and Telecommunications* (1): 1–12.

Deibert, Ronald J. 2020. *Reset: Reclaiming the Internet for Civil Society*. Toronto: House of Anansi Press.

Ding, Xiaodong. 2014. "Exploring Legal Standards for Anti-Discrimination and Equality Protection: An Analysis Based on the 'Disparate Impact Standard'." *Peking University Law Journal* (4): 1080–96.

Dworkin, Ronald. 1986. *Law's Empire*. Cambridge: Belknap Press.

Dworkin, Ronald. 2013. *Taking Rights Seriously*. London: A&C Black.

Elyamany, N. 2024. "The De-legitimation of Machine Learning Algorithms (MLAs) in "The Social Dilemma"(2020): A Post-digital Cognitive-Stylistic Approach." *International Journal of Legal Discourse* 9 (1): 59–92.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Feller, Avi, Emma Pierson, Sam Corbett-Davies, and Sharad Goel. 2016. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not that Clear." *Washington Post* 17.

Floridi, Luciano. 2023. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.

Foucault, Michel. 1975. *Discipline and Punish.* Translated by A. Sheridan. Paris: Gallimard.

Gao, Yifei. 2019. "Protection of Rights of Digital Vulnerable Group in Intelligent Society." *Jianghai Academic Journal* (5): 163–9.

Garland, David. 2002. *The Culture of Control: Crime and Social Order in Contemporary Society*. Oxford: Oxford University Press.

Gottfried, Wilhelm Leibniz. 2013. *Logico-Philosophical Puzzles in the Law: Philosophical Questions and Perplexing Cases in the Law*. Dordrecht: Springer Science Business Media.

Habermas, Jürgen. 1996. *Contributions to a Discourse Theory of Law and Democracy*. Translated by William Rehg. Cambridge: Polity Press.

Harcourt, Bernard E. 2006. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago: The University of Chicago Press.

Hart, Herbert Lionel Adolphus. 2012. *The Concept of Law*, 3rd ed. Oxford: Oxford University Press.

Holmes, Oliver Wendell. 1946. "Learning and Science." *North Carolina Law Review* 24: 102.

Holmes, Oliver Wendell. 1997. "The Path of the Law." *Harvard Law Review* 110 (5): 991–1009.

Ji, Weidong. 2007. "Dialectical Analysis of Criminal Punishment Imposition through Software." *Tribune of Political Science and Law* (1): 124–8.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–99.

Kant, Immanuel. 1998. *Groundwork of the Metaphysics of Morals*. Cambridge: Cambridge University Press.

Kelsen, Hans. 2017. *General Theory of Law and State*. London: Routledge.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and C. R. Sunstein. 2018. “Discrimination in the Age of Algorithms.” *Journal of Legal Analysis* 10: 113–74.

Lei, Lei. 2015. “Legal Methods, Legal Certainty and the Rule of Law.” *The Jurist* (4): 1–19.

Li, Cheng. 2021. “Legal Governance of Artificial Intelligence Discrimination.” *China Legal Science* (2): 127–47.

Li, Xunhu. 2021. “Inclusive Regulation of Artificial Intelligence in Criminal Justice.” *Social Sciences in China* (2): 42–62.

Liu, Wenyan. 2021. “Survey on Fairness in Trustworthy Machine Learning.” *Journal of Software* (5): 1404–26.

Meng, Xiaofeng, and Xiang Ci. 2013. “Big Data Management: Concepts,Techniques and Challenges.” *Journal of Computer Research and Development* (1): 146–69.

Molnar, Christoph. 2020. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Vancouver: Leanpub.

Noble, S. U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

O’Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group.

Paul, R. Cohen & Edward A. Feigenbaum. 1982. *The Handbook of AI*, Vol. III. Stanford: Heuris Tech Press; Los Altos, California: William Kaufmann.

Pauline, Kim. 2016. “Data-driven Discrimination at Work.” *William and Mary Law Review* 58: 857.

Posner, Richard A. 2014. *Economic Analysis of Law*. New York: Aspen Publishers.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Rao, U., and V. Nair. 2019. “Aadhaar: Governing with Biometrics.” *South Asia: Journal of South Asian Studies* 42 (3): 469–81.

Sandvig, C. 2022. “The Regulation of Algorithmic Decision-Making in the Courts: A Global Comparison.” *Law and Technology Review* 18 (4): 335–50.

Schwemer, S. F., L. Tomada, and T. Pasini. 2021. “Legal Ai Systems in the Eu’s Proposed Artificial Intelligence Act.” In *Proceedings of the Second International Workshop on AI and Intelligent Assistance for Legal Professionals in the Digital Workplace (Legal AIIA 2021)*. São Paulo: ICAIL 2021.

Song, Lijue, and Changshan Ma. 2022. “Identifying the Fourth Generation of Human Rights in Digital Era.” *International Journal of Legal Discourse* 7 (1): 83–111.

Sun, Jianli. 2019. “Research on Regulations of Algorithm Automatic Decision Risk.” *Research on Rule of Law* (4): 108–17.

Surden, Harry. 2011. “The Variable Determinacy Thesis.” *Science and Technology Law Review* (12): 1–91.

Suresh, Harini and John V. Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002.2 (8): 73.

Susskind, Richard. 2008. *The End of Lawyers? Rethinking the Nature of Legal Services*. Oxford: Oxford University Press.

Susskind, Richard. 2023. *Tomorrow’s Lawyers: An Introduction to Your Future*. Oxford: Oxford University Press.

Sweeney, Latanya. 2013. “Discrimination in Online Ad Delivery.” *Communications of the ACM* 56 (5): 44–54.

Turing, Alan Mathison. 1936. “On Computable Numbers, with an Application to the Entscheidungsproblem.” *Journal of Mathematics* 58: 345–63.

van Dijk, Jan A. G. M. 2006. “Digital divide research, achievements and shortcomings.” *Poetics* 34 (4–5): 221–35.

Weber, Max. 2019. *Economy and Society: A New Translation*. Translated by Keith Tribe. Cambridge: Harvard University Press.

Wang, Juan, Ci Linlin, and Kangze Yao. 2005. “A Survey of Feature Selection.” *Computer Engineering & Science* (12): 72–5.

Wirth, Rüdiger, and Jochen Hipp. 2000. “CRISP-DM: Towards a Standard Process Model for Data Mining.” In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, Vol. 1, 33–4.

Yeung, Karen. 2018. “Algorithmic Regulation: A Critical Interrogation.” *Regulation & Governance* 12 (4): 505–23.

Yu, Xingzhong. 2016. “When Law Meets AI.” *Legal Daily* (3): 76.

Zarsky, Tal. 2021. “The Problem of “Algorithmic Discrimination” in Judicial Artificial Intelligence.” *Yale Journal on Regulation* 38 (2): 512–44.

Zarsky, Tal. 2016. “The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making.” *Science, Technology & Human Values* 41 (1): 118–32.

Zhang, Su. 2014. “Empirical Analysis of the Application of Arrest Measures for Floating Population Crimes —Taking the Arrest Practices in District A of Beijing as the Main Sample.” *Journal of People’s Public Security University of China(Social Sciences Edition)* (3): 72–4.

Zhang, Wenxian. 2019. “Human Rights Jurisprudence in the New Era.” *Human Rights* (3): 12–27.

Zuboff, Shoshana. 2023. *The Age of Surveillance Capitalism. Social Theory Re-Wired*, 203–13. London: Routledge.

Bionotes

Landuo Dou
Guanghua Law School, Zhejiang University, Hangzhou, China
Law and Technology Institute, Renmin University, Beijing, China
landuo_dou@zju.edu.cn

Landuo Dou is a Research Assistant of the Digital Law Research Institute of Zhejiang University, Law and Technology Institute of Renmin University. Her research interestes are digital law, data governance, artificial intelligence law and philosophy of law.

Xiaodong Dou
Guanghua Law School, Zhejiang University, Hangzhou, China
Institute of New Era Fengqiao Experience, Zhejiang University, Hangzhou, China
douxd@zju.edu.cn

Xiaodong Dou is a professor of the Law School at Zhejiang University, director of the Environmental Resources and Energy Law Research Center at the Law School of Zhejiang University, executive vice dean of the Institute of New Era Fengqiao Experience at Zhejiang University, and president of the Maritime Economy Rule of Law Research Association of the Zhejiang Law Society. His research interests and publications cover a wide range of areas, including digital law and governance, environmental policy and law.