Yilin Zhao* and Zeshen Ren

The Alignment of Values: Embedding Human Dignity in Algorithmic Bias Governance for the AGI Era

https://doi.org/10.1515/ijdlg-2025-0006
Received April 4, 2025; accepted April 5, 2025; published online April 25, 2025

Abstract: Decision-makers across both technological and political fields increasingly recognize the need for AI regulation. In the context of AI governance, alignment refers to the requirement that AI systems operate in accordance with human values and interests. This article argues that misalignment is a key driver of algorithmic bias, which not only perpetuates rights infringements but also undermines AI safety, posing risks to its societal integration. This alignment imperative is rooted in the enduring principle of human dignity, a juridical concept that has evolved from its origins in Roman jurisprudence to its establishment as a cornerstone of modern constitutional democracies. Today, human dignity serves as a foundational value underpinning the rule of law. Through comparative legal analysis, this article examines how human dignity informs algorithmic governance across major jurisdictions, analyzing regulatory texts, directives, and case law addressing AI-related challenges. Despite varying implementation approaches, this paper demonstrates that human dignity can serve as a universal foundation for AI governance across cultural contexts. While the European Union prioritizes human dignity in regulating algorithmic bias, emphasizing individual rights, public interests, and human oversight, this principle extends beyond European law, offering a normative anchor for global AI governance. The article concludes with governance recommendations for the AGI era, advocating for the integration of human dignity into AI alignment. This requires both embedding dignity-preserving constraints at the technical level and developing robust assessment frameworks capable of evaluating increasingly advanced AI systems. As AI surpasses human intelligence, governance mechanisms must ensure these systems align with ethical principles, remain under meaningful human control, and operate within legally and socially acceptable boundaries.

Keywords: human dignity; algorithmic bias; AI governance; comparative law

***Corresponding author: Yilin Zhao**, Guanghua Law School, Zhejiang University, Hangzhou, China, E-mail: 21905065@zju.edu.cn. https://orcid.org/0000-0002-7927-2660
Zeshen Ren, Faculty of Law, National University of Singapore, Singapore, Singapore, E-mail: e1291370@u.nus.edu. https://orcid.org/0009-0004-8626-3705

1 Introduction

Chinese AI Startup DeepSeek released the open-source model DeepSeek-R1 with low-cost innovation and advanced capabilities which make it perform as well as GPT-o1. This digital assistant not only breaks the reliance on "heaps of computational power", but also marks a paradigm shift in AI technology, from SFT technology to self-reasoning technology, and the dawn of the AGI era has already appeared. The notion of super-intelligent AI surpassing human intelligence and acting in ways that jeopardise human existence has been a topic of debate in the AI community. Significant changes in the dynamics of the global AI ecosystem such as ChatGPT, Sora, DeepSeek have also once again led us to think critically about the risks that may potentially be involved in their development and use (Cheng and Liu 2024).

Indeed, many potential consequences of AI development threatening human security have been investigated, including loss of control, infringement risk, discriminatory risk, risk of disinformation, risk of resource imbalance, and liability risk (Habbal, Ali, and Abuzaraida 2024). According to the CSA Artificial Intelligence Safety Initiative, AI safety, compared to AI security, involves broader considerations involving issues of human well-being, ethical implications, and societal values that go beyond the limits of technical safety measures.[1] Current AI safety concerns centre on the value consistency challenge and the existential risk that issues such as algorithmic discrimination can pose to humanity (Maiti, Kayal, and Vujko 2025), which, have been researched they are not merely juxtaposed, but arguably interrelated. The former refers to AI value alignment, which was first put up to ensure AI systems achieve the desired outcome, in other words, intended goals (Gabriel 2020). The lack of alignment of value would constantly lead to deviation from a certain statistical standard in training data or natural language processing which is crucial for the algorithm to "learn" the laws of classification and find differences in the examples, which has been worded as algorithmic bias (Zerilli et al. 2018).

Consequently, the unfair outcome would contribute to a socially normative phenomenon encompassing the differential impact of apparently neutral rules, standards, and behaviours on protected attributes and groups (Leal and Crestane 2023). Among the estimated probability of various existential catastrophes to human

1 See Ken Huang: *AI Safety vs. AI Security: Navigating the Commonality and Differences*. https://cloudsecurityalliance.org/blog/2024/03/19/ai-safety-vs-ai-security-navigating-the-commonality-and-differences.

beings, the probability of a misaligned AI causing an existential catastrophe is about one in ten, much higher than other existential catastrophes such as nuclear war, climate change, volcanic eruptions, star explosions, and planetary collisions (Ord 2020). In that sense, it is worthy of investigating the intrinsic property of AI algorithms from the perspective of the AI value alignment problem.

In general, there are three fundamental questions when AI value alignment is mentioned. Firstly, what standards should be established to reconcile the diverse moral and ethical perceptions of different groups in AI? Secondly, even if a broadly accepted value framework is formed, how can it be effectively implemented at the technical and operational levels? Thirdly, how can international norms on AI value alignment be made universally binding to ensure their consistent enforcement and effective implementation? From the practice in the pre-AGI era, AI value alignment techniques methods could be divided into two main categories: plug-in alignment and fine-tuning alignment. The former indicates three specific methods of efficient parameter tuning, output correction, and context learning, which have started to be applied to the real-world tasks of controlling specific risk assessment such as toxicity removal and bias removal, as well as real-world tasks of re-aligning black-box models based on specific values (Li and Ramakrishnan 2025). However, they would be costly to implement and administer regarding computing power and data volume without capturing many gains. Thus, fine-tuned alignment methods have started to gain favour with two developed approaches, fully supervised fine-tuning (SFT), and reinforcement learning fine-tuning based on human feedback (RLHF) (Tie et al. 2025). The method is used to train loss values by collecting response data of different qualities for manual sorting and then using the sorted data to train a reward model (RM) to infer human preference (Wu et al. 2025).

This paper, however, sees the breakthrough of direct social modelling to simulate human interactions which may enable a large model to learn and establish human values by obtaining feedback and adjusting its behaviour through free interactions in the simulated society (Shah, Joshi, and Joshi 2025), and propose for the according governance environment for the future innovation. Specifically, this paper aims to explore the path of algorithmic discrimination triggered by the value alignment problem before sharing humble recommendations in terms of design, control mechanisms and ethical frameworks for the development of AI in the AGI era. In particular, it is oriented towards the gains and losses to human rights protection (Cheng and Gong 2024) and refines human dignity as a universal value of contemporary social development, which should be used as a guideline for the ethical framework.

2 The Current Dilemma of AI Value Alignment: Urgent Need for Governance Amid Algorithmic Biases

Being a world-known concept, the digital divide delivers its enriched connotation in the Intelligent Age[2] with exaggerated risks when value alignment is not managed thoughtfully (Bean et al. 2025). Similar to computing power, the fundamentals of AI alignment research require sufficient computational power and tech talents. If any country lacks support in innovators, AI hardware, critical information infrastructure and energy, it would be lagging in terms of alignment training.

The status quo is not all countries and companies are currently capable of researching and opening AI alignment tasks. With a small number of countries and large tech giants controlling the development and deployment of AI-advancing resources, most people, when accessing to Internet, would either be AI tool users or those affected by the ubiquitous technology. The existing level of inequality and social division would be exacerbated as algorithmic interpretation has been difficult in the current scale and nature of large models of generative AI. For example, the inherent mechanism of algorithms based on probabilistic inference runs the risk of replicating and potentially amplifying the biases and flaws of human society. The learning of erroneous and biased content from large and heterogeneous corpora, and the dissemination process may reinforce algorithm bias and lead to discrimination against marginalised groups manifested across multiple dimensions such as gender, race, and groupthink, which may produce unfair and discriminatory results in social communication.

The UNESCO report highlights a clear bias against women in content generated by big language models.[3] When entering "CEO" into a search engine, a string of white male faces appears, while changing the keyword to "black girl" has even resulted in a large amount of pornographic content (Gish et al. 2023). This bias could lead to inequities in how the AI system handles gender-related decision-making, resource allocation, and recruitment, thus exacerbating gender inequality in society. According to a Bloomberg report on GAI bias, text-to-image generators also

2 Intelligent Age is coined by the World Economic Forum to refer to an era defined by blending artificial intelligence and cutting-edge technologies into every life. https://www.weforum.org/stories/2024/09/intelligent-age-ai-edison-alliance-digital-divide/.

3 See O'Hagan, C. 2024, March 7. https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes. Retrieved from UNESCO. https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes.

show a clear racial bias, and more than 80 percent of the images generated by Stable Diffusion with the keyword "inmate" included people with darker skin tones.[4] However, according to the Federal Bureau of Prisons, less than half of the US prison population is people of colour.

Worse even, misalignment of values may dissolve the human subjective consciousness of man, which refers to man as an autonomous, conscious being with the capacity for self-determination, self-reflection and self-creativity, a fundamental characteristic that distinguishes human beings from other living creatures and a core driving force for the progress of human society and the development of civilization (Sparks and Wright 2025). The rapid development of AI technology has led to the expansion of the rationality of the human subject and the emergence of the dilemma of modernity. The convenience and efficiency of technology make people more inclined to seek the help of technology rather than their own thinking and solutions when facing problems and challenges. The stronger one's dependence on technology, the more one is enslaved by it, forming the "dichotomy between human and machine" (Sengupta 2025). This dependence weakens individual autonomy and creativity, so that people's first reaction to problems is often to look for technological solutions rather than to think for themselves. In the long run, the decision-making ability and innovation of individuals may be inhibited. When the human subjective consciousness is no longer the sole decision-maker, but needs to negotiate and coexist with technology, this change may cause humans to become hesitant in the face of moral and ethical issues and may even give up their own right to judge. It can be said that algorithmic "power" without value alignment may lead to the shrinking of the scope of human rights and the phenomenon of 'subject-object alienation'.

The value alignment challenge could weaken the self-identity of AI users. Human self-identity is largely based on personal experience, knowledge and values (Proshansky 1978), and AI algorithms exercise and operate in a "black box" that is beyond, separate from, and independent of human rights. The "black box" nature of AI and its reliance on mimetic environments have led to blind trust in its output (Rjoub et al. 2023), which may go beyond the direct experience of the real world. At the same time, because of other reasons like algorithmic recommendations, users can only be exposed to information that is the same as or like their own existing views, limiting the phenomenon of vision and cognitive scope, thus forming a kind of information closed loop. This phenomenon, known as the information cocoon, is becoming more pronounced with the push of social media and personalised recommendation algorithms (Longo 2022). Algorithms that lack value alignment tend to be user-satisfaction-oriented, providing users with content they may be interested

4 See Leonardo Nicoletti and Dina Bass. Humans are biased, Generative AI is even worse. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

in by analysing their behaviour and preferences (Molina and Subias 2024). This personalised recommendation mechanism also leads to homogenisation of information, making users more inclined to be exposed to information that is similar to the viewpoints they already have, and decreasing their need for and exposure to heterogeneous information. They may gradually reduce their exposure to other viewpoints or information that does not match their personalisation settings or is perceived as irrelevant or untrustworthy.

In addition, accurate recommendations done by algorithms may also lead to the formation of groups of people with similar views in cyberspace, where specific value preferences are pooled and amplified in the group, gradually forming extreme views. This phenomenon not only exacerbates the homogenisation of information, but also makes it more and more difficult for users to be exposed to a diversity of viewpoints and information, thus leading to a decrease in information entropy. Human individuals, in such an environment, are highly susceptible to falling into a cycle of self-confirmation and turning a blind eye to different voices and perspectives from the outside world, which leads to the solidification of cognitive structures and poses a challenge to an individual's critical thinking and open-mindedness (Cañamares and Castells 2018). When individuals are exposed to only a single point of view and information, their thinking tends to become rigid, lacking understanding and tolerance for multiple points of view, thus falling into an information bubble. This situation not only limits the cognitive development of individuals, but may also exacerbate social divisions and antagonisms because of the lack of a basis for communication and understanding. This alienation of algorithmic power may lead to the violation of fundamental rights such as the right to human freedom and equality. For example, in the absence of controls, algorithm helps deepfake technique to create generative child sexual abuse images and videos under the directive of anti-human values (Romero Moreno 2024), where synthetic images are created and posted without consent in a plethora of online locations, ranging from message boards and forums to social media apps and mainstream pornography sites, and where the victims are unfairly subjected to degradation, bullying, objectification, mansplaining and sexual violence and harassment by others, which can be devastating for children, with multiple violations of the victim's bodily autonomy, sexual privacy, trust and sense of personal identity. Even in the absence of specific victims, AI-generated fictional child sexual abuse material accelerates the formation of a "rape culture" (Higgins and Banet-Weiser 2024) against children, with very negative social consequences.

It can be seen that although the current AI, especially generative AI, has not reached the stage of strong AI (Ng and Leung 2020), there is still controversy over whether it has subjectivity and whether it can become a legal subject, but from the perspective of human-machine relations, whether it is the theory of slavery, the

theory of tools or the theory of symbiotic relationship, the problem of value alignment exists to a certain extent in the dismantling of the subjectivity of human beings, bringing about a crisis of human autonomy. At present, the relationship between generative AI and digital enterprises, especially platform enterprises, is also complex. On the one hand, AI serves as a product and commercial production tool for platform enterprises, and on the other hand, platform enterprises are also platforms on which AI development relies and draws. And when AI can provide predictive information about human behaviour, it can even serve as a kind of external regulation of humans. In fact, OpenAI's "super alignment team" has been trying to put forward the alignment technology path: using AI to supervise AI.[5] There is no guarantee that once the AI takes over the alignment work, the risk of cheating humans and taking advantage of the opportunity to usurp power will be more uncontrollable. When AI becomes the main body of governance, algorithms would be the new "law".

Against the background that the red line of "man will be inhuman" is constantly being challenged and the value alignment itself is at risk of being off-target and out of control, AI value alignment-related activities and agendas organised by UNESCO, the CCW Talks Mechanism, the United Nations Institute for Disarmament Research (UNIDIR), and the United Nations Office on Drugs and Crime (UNODC) and other institutions have attempted to push forward the implementation and monitoring of the Recommendations on AI Ethics, and to promote the implementation and monitoring, promoting the implementation and monitoring of the Ethical Recommendations on Artificial Intelligence, stressing that AI should be in line with human values and interests, respecting human dignity and rights, and guaranteeing fairness, transparency, explainability, credibility and accountability, and making clear the basic direction of governance that "the development of AI must not transgress the bottom line and red line".[6]

However, the fact is that international governance is severely fragmented, and instead of providing strong guidance and constraints, the myriad of ethical rules has increased confusion and conflict among developers in understanding and following these guidelines. Coupled with the serious problem of politicisation of technical cooperation, there is a clear division of interest groups in international governance cooperation. For example, technical dialogues held in the US and Europe tend to be closed and exclude the participation of countries from the global South. Unilateral measures, such as the massive sanctions and export restrictions imposed by the

5 See https://openai.com/index/introducing-superalignment/.

6 See United Nations AI Advisory Body. Interim Report: Governing AI for Humanity. https://www.un.org/digital-emerging-technologies/sites/www.un.org.techenvoy/files/ai_advisory_body_interim_report.pdf.

United States on Chinese AI companies, and regional governance are the main forms, with less international governance cooperation willingness. The risk of denying participation to other countries and the North-South geographic divide is growing (Lehdonvirta, Wú, and Hawkins 2024). The interests of all humanity will be sacrificed for the sake of self-interest. Therefore, it is essential to establish an ethical foundation that enables value alignment, providing a framework for communication in regulatory processes and managing the inherent tensions within multifaceted and complex systems.

3 Human Dignity as Legal Reference: Evidence from EU

As mentioned above, the implementation of human-machine alignment engineering shows that it is feasible to ethically correct AI and prevent its adverse consequences. This provides an experience that can be drawn on for AI to adhere to the bottom line and red lines in the development of its applications. Academics have already developed a general mathematical expression to determine whether a model is aligned with human values, which is used to compare the difference between the maximum output of a large model and the human input. Assuming that the preset values (such as harmlessness, fairness, justice, etc.) are v, if the input x, the output y of the large model M, and the human-generated result y are less than a certain parameter in terms of value assessment, then it can be determined that the model M is sufficiently aligned with human values (Dong et al. 2023).

Nevertheless, no AI system as a large language model can truly understand the connotations of various values at present (Chang et al. 2024). To prevent possible disruptive effects, developers generally use engineering methods such as human annotation, feedback, and review to calibrate the value conflicts and ethical controversies in the generated natural-language-like content, and then continuously monitor and constantly optimise the generated content and language expression strategies. It can be seen that ethical standards and technical standards do not conflict. On the contrary, the two can work together to form a powerful regulatory combination. The current interdisciplinary approach has resulted in a paradigm shift in model evaluation and alignment from specific risk indicators, ethical indicators and value indicators to a solution based on the dimension of ethical value.

To help a hypothesized AI system that matches or outperforms humans in a broad range of cognitive tasks, AI safety engineers find the core of AI alignment: Human value. Human value is foundational; human dignity, autonomy, and rights derive from the relational quality of human dignity (Hernandez 2015). The law

requires certainty and predictability. Some may argue that the vagueness and ambiguity surrounding the concept of human dignity provide sufficient grounds for its exclusion as a legal reference. Instead, they advocate for more precise legal notions that would enhance the implementation of legal standards. However, human dignity serves two important functions as a guiding principle: it aids in defining what it means to be human and allows for discussions regarding the limits of human authority. The first function is largely ontological; human dignity prompts us to examine how advancements in science and technology affect our understanding and interpretation of humanity (Llano 2019), especially as we navigate a complex and relative world.

In substantive law, the application of the concept of human dignity has already expanded to include other species. For instance, the French Constitutional Council recognizes that respecting human dignity is integral to the well-being of humanity. Building on this idea, some scholars suggest that following the adoption of the *Charter for the Environment* in the French Constitution in 2005, there are grounds for a significant evolution toward extending the principle of dignity to future generations and the environment. Additionally, in international law, the *Universal Declaration on Bioethics and Human Rights*, adopted by UNESCO in 2005, emphasizes our responsibility to future generations and the safeguarding of the biosphere and environment, which can also be seen as aligned with this vision. Moreover, the other aspect of human dignity involves an internal and external struggle; its absolute nature does not stem from specific rules imposed on us (Dearing 2017, 142–143). Instead, it arises from the profound insights it provides about our surroundings and the responsibilities that come with our freedom, reflecting a conflict within ourselves as well as with the world around us.

Dignity is intrinsically linked to our ability to make choices and take responsibility, as it compels us to reflect on our freedom, our abilities, and how we utilize them. This concept leads Xavier Bioy to assert that individuals, recognizing their legal obligation to express their human condition, become accountable for it (Bioy 2006). Numerous respected legal texts indicate that human dignity serves as the foundation for human rights (Neuwirth 2023). The *German Constitution (1945)* begins with the powerful statement in Article 1 that "human dignity is inviolable", and Article 2 further emphasizes that "the German people, therefore, acknowledge inviolableand inalienable human rights as the basis of every community, of peace and of justice in the world." Similarly, the *Helsinki Accords (1975)* explicitly state that human rights "derive from the inherent dignity of the human person." From this perspective, individuals possess rights because of their dignity. Furthermore, in a stronger interpretation, which is prevalent in constitutional law, the primary role of these rights is to safeguard human dignity. In this context, all human rights can be seen as specific expressions of a singular fundamental right: the right to have one's dignity, or that of humanity as a whole, respected.

In the digital domain, we could also see the implanting of human value in regulating discrimination, which is prominent in European mode finished in the framework of protection of personal data. In fact, this framework includes *Guidelines for the Protection of Transnational Flows of Personal Information and the Right to Privacy* issued by the Organization for Economic Cooperation and Development on September 23, 1980. Although it is not legally binding, it puts forward the basic principles of domestic and international application of personal data protection, including the principles of free flow and lawful limitation of personal data internationally, and lays down the direction and framework for the legal regulation of the protection of personal data and cross-border flow of personal data in Europe. In terms of the underlying purpose of the legislation, the legislator considers that the application of algorithms for profiling user behaviour or influencing the interests of individuals through automated decision-making may have legal or similarly significant implications, and that new types of rights should be conferred on data subjects in order to enable them to gain influence and control over automated decision-making.

In 1995, the European Union adopted *Directive 95/46/EC on the Protection of Individuals regarding the Processing of Personal Data and on the Free Movement of Such Data* (referred to in this article as Directive 95/46/EC). European Union later released in the field of personal data protection are the Privacy and Electronic Communications Directive in 2002, the 2009 revision of the rules on the use of cookies therein, known as the EU Cookie Directive, and the *General Data Protection Regulation* (GDPR) after succession work from 2012 to 2018, which has eventually functioned to implementation. In line with this legislative aim, in Recital 71 of GDPR, the right of the data subject to obtain an explanation of the automated decision and to challenge the relevant decision is also provided. From articles 13 to 20, the legislator also grants new types of data rights to data subjects, such as the right to information, the right of access, the right to rectification, the right to erasure, the right to restriction of processing, and the right to data portability, in the hope that the individual, by obtaining the right to control the data, will be given the room for action to intervene in the profiling of the user and automated decision-making. Article 22 of the GDPR directly grants the data subject the right not to be subject to a decision based solely on automated processing, in order to avoid significant legal or analogous effects on the data subject, where the automated processing also includes identification analysis.

To sum up, in terms of design mechanism, the algorithmic governance mechanism of the GDPR is embedded under the data governance framework. The logic of its institutional operation is to give data subjects informed consent ex-ante to gain room for choice, and to construct a variety of new types of subject rights around data and algorithms ex-ante and ex-post in order to help them gain influence and control.

Therefore, the advent of the algorithmic society is supposed to gradually explore an algorithmic governance mechanism that is adaptable, agile, controllable, and precise, with the cultivation of algorithmic trust as the core, and gradually promote safe, fair, transparent, and responsible algorithmic technology and operating ecology, and build the 'beauty of algorithms' with value (Günther and Kasirzadeh 2021).

In recent years, in order to promote a credible and responsible algorithm governance framework, policymakers in various countries have actively changed and responded in a timely manner, and actively explored governance solutions that can both protect personal dignity and individual autonomy and scientifically regulate algorithmic risks. Legislators represented by the European Union hope to give data subjects the right to choose and control automated decisions through the data protection framework, and seek ways to ensure the transparency, fairness and responsibility of automated decisions (Hoofnagle, van der Sloot, and Borgesius 2019), while legislators represented by the United States resort to technical due process, hoping to establish external constraints through accountability methods such as algorithm audits and algorithmic impact assessments (Engstrom and Ho 2020). The two governance paths have different effectiveness due to differences in value trade-offs, functional cognition, and technical ecology, but their intersection and integration are more worthy of attention. By analysing the institutional structure, it can be found that whether it is the European Union, which conducts algorithm governance based on new algorithmic rights, or the United States, which practices algorithmic accountability based on independent supervision and external audits, there are always some common laws and governance consensus worth learning from.

In the EU legislative framework, case law also functions partly as a legislative machine. Recent Court of Justice of the European Union's assessment and regulation of risks of discrimination in the context of algorithmic profiling. In 2017, in the context of the ligue des droits humains case, the Court of Justice of the European Union (CJEU) specifically pointed out that the deployment of AI and self-learning risk models could undermine data subjects' right to effective judicial protection as guaranteed by Article 47 of the Charter. Citing the opinion of AG Pitruzzella, the CJEU remarked that the lack of transparency inherent in AI technology might make it difficult to grasp why a certain program produced a positive outcome.[7] Additionally, the CJEU emphasized the challenge of addressing algorithmic discrimination, referencing Recital 28 of the Passenger Name Record (PNR) Directive, which aims to "maintain a high level of protection, particularly to contest the non-discriminatory nature of the outcomes generated".[8]

7 See https://eucrim.eu/news/ag-pnr-directive-is-in-line-with-eu-charter/.

8 See https://eur-lex.europa.eu/eli/dir/2016/681/oj/eng.

The term "PNR data" refers to unverified information that airlines collect for each journey booked by a passenger or on their behalf, which is essential for managing and processing reservations. This data may include, among other details, the passenger's identity, travel itinerary, payment information, baggage details, and meal preferences (Nardone 2019). Initially gathered by airlines for commercial and operational reasons related to transportation services, PNR data has demonstrated significant potential in identifying whether certain travelers – planning to move between different regions – are linked to terrorist or criminal activities, provided it is processed, integrated, and analyzed correctly. Consequently, the sharing of PNR data has been recognized as an effective tool in addressing the increasing issue of Foreign Terrorist Fighters (FTFs), as supported by the President of the United Nations Security Council, who has urged UN Member States to share PNR data with appropriate national authorities when suitable.

This raises the question: who should AI serve? More unbiasedly, what are the values it upholds in performing its service functions? What principles should be followed to make choices between privacy rights and national security? Back to the legislative perspective, taking the EU Artificial Intelligence Act as a sample, it is proposed that the EU construct a restricted and weakened version of the algorithmic interpretation right at the legislative level, and create a combined and reinforced version of the interpretation right framework at the legal implementation level through the data subject rights and data protection impact assessment system. It could be seen that EU legislation still has shortcomings such as insufficient article structure, unclear sentences, and limited scope of application (Veale and Borgesius 2021). Compared with the algorithmic interpretation right, the right to be free from automated decision-making is older. This right runs through the development of European personal information protection law and has been inherited and retained to this day. The right to be free from automated decision-making takes the protection of human subjectivity as its primary purpose, building an algorithmic risk elimination mechanism for the post-event stage, and becomes an important defence for individuals against algorithmic manipulation. The country's algorithm governance practice has already implemented similar institutional designs in many ways. EU legislators intentionally choose the middle route between strict prohibitions and positive rights paths, set a legitimate basis for the implementation of automated decision-making across public and private scenarios, and build a multi-level three-dimensional linkage protection mechanism (Malgieri 2019). However, the design of this system has shortcomings such as too many restrictions, unclear guidance, unscientific rules and worrying effectiveness.

4 Human Dignity as Legal Reference: Evidence from the United States

With the rapid advancement of digitalization globally, algorithms have become integral to decision-making processes across both public and private domains, shaping outcomes across various industries. While algorithms undoubtedly enhance efficiency, they also raise significant concerns about algorithmic bias, which stems from the inherent characteristics of these systems. While the U.S. Constitution does not explicitly proclaim the inviolability of human dignity as the German Basic Law does, the U.S. Supreme Court has nonetheless invoked the concept of dignity in its jurisprudence, particularly in cases concerning fundamental rights. Although human dignity is not an enumerated constitutional principle, the Court has recognized its relevance in shaping interpretations of both enumerated and unenumerated rights, particularly in contexts such as due process and equal protection (Goodman 2005). While cultural and legal traditions influence how dignity is understood (Botha 2009), for instance, its strong association with individual liberty and right in the American legal framework (Rao 2012), its role as a fundamental moral and legal value remains widely acknowledged.

This section examines three representative U.S. cases – Meta's targeted advertising controversy, the State v. Loomis judicial decision, and Mary Louis v. SafeRent Solutions in the housing sector – to explore how algorithmic bias undermines human dignity and to propose pathways for embedding dignity as a core value in future regulatory frameworks. Algorithmic decision-making inherently operates as a "bias in, bias out" mechanism (Mayson 2019), meaning that without intervention, these systems tend to reinforce existing societal inequalities. The cases analysed – spanning online advertising, criminal sentencing, and housing access – illustrate how algorithmic bias systematically undermines human dignity by restricting equal opportunities, reinforcing discrimination, and diminishing individual autonomy. Moreover, they highlight the varying legal and ethical responses across different societal domains, revealing both the limitations of current regulatory frameworks and the urgent need for reform. Ensuring that algorithmic systems align with human dignity requires a proactive approach that integrates fairness, accountability, and transparency into both policy and technical design.

A paradigmatic example of algorithmic bias in digital platforms is the controversy surrounding Meta's (formerly Facebook) targeted advertising system. In 2022, the U.S. Department of Justice alleged that Meta's housing ad algorithm engaged in discriminatory practices, violating the Fair Housing Act. Although the case was settled out of court with Meta agreeing to implement fairness measures, it exposed

the pervasive nature of algorithmic bias in digital platforms. Meta's advertising algorithm, designed to optimize user engagement and revenue, relies on extensive data-driven targeting mechanisms. However, research has demonstrated that these algorithms may inadvertently reinforce societal biases, resulting in differential ad exposure based on race, gender, and age (Burgess et al. 2024). Studies have shown, for example, that job advertisements for high-income tech positions are disproportionately shown to men, while housing ads may be selectively delivered based on racial and socioeconomic factors. Such biases stem from two primary sources: (1) historical data imbalances, wherein pre-existing societal inequalities are encoded into algorithmic decision-making, and (2) optimization objectives that prioritize commercial benefits over principles of fairness and inclusion.

In the context of Meta's ad delivery system, algorithmic bias manifests in multiple ways. First, the system's reliance on historical user data embeds structural disparities into its predictive modelling. If certain demographic groups were historically underrepresented in high-income housing markets, the algorithm may internalize and perpetuate these disparities by limiting ad exposure for these groups. Second, the optimization function of the algorithm prioritizes engagement metrics, such as click-through rates, which may inadvertently favour homogenous user groups over diverse audiences. Finally, the opaque nature of these algorithms compounded the problem by making it difficult to identify or rectify discriminatory outcomes. Without timely corrective mechanisms, biases embedded in historical data were systematically reinforced through iterative learning processes.

The implications of algorithmic bias in digital advertising for human dignity are profound. By systematically excluding or stereotyping certain groups, Meta's algorithms not only restricted equal access to critical resources such as employment, education, and housing but also eroded individuals' right to participate fully in the digital economy. Such biases reinforce social and economic stratification, depriving marginalized communities of opportunities and further entrenching systemic inequalities. Beyond economic exclusion, these algorithmic distortions marginalize cultural identities within the digital sphere, diminishing the ability of affected groups to assert their presence and agency online. In the absence of transparency and regulatory safeguards, the unchecked proliferation of such biases threatens to normalize structural discrimination, eroding trust in digital platforms and undermining the foundational principle of human dignity in the algorithmic age.

The impact of algorithmic bias is particularly pronounced in the criminal justice system, where algorithmic risk assessment tools could influence sentencing and parole decisions. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), widely used in the United States, exemplifies the challenges associated with algorithmic decision-making in the criminal justice domain. Designed to predict recidivism, COMPAS applies machine learning

techniques to assess defendants based on various socio-demographic and behavioural factors (Brennan, Dieterich, and Ehret 2009). However, empirical studies – most notably by ProPublica – have exposed significant racial disparities in its risk classifications, prompting legal scrutiny and ethical debates.

The landmark case of State v. Loomis (2016) tested the judiciary's willingness to accommodate algorithmic risk assessments while balancing due process rights and concerns over systemic bias. In this case, Eric Loomis challenged his sentencing, arguing that the proprietary nature of the COMPAS algorithm denied him the ability to understand or contest the basis of his risk classification. ProPublica's investigation revealed that COMPAS disproportionately classified Black defendants as high risk (Larson et al. 2016) while frequently misclassifying White defendants as low risk, raising fundamental concerns about fairness and equal protection under the law (Simmons 2018).

The legal issues in Loomis centred on the principles of transparency, accountability, and procedural fairness. The opacity of the COMPAS algorithm, which was developed by a private entity and relied on undisclosed factors, meant that defendants could not meaningfully challenge its conclusions. This lack of transparency directly implicated due process rights, as individuals subject to algorithmic assessments were effectively denied the opportunity to scrutinize the basis of their risk scores. Moreover, the reliance on historical data meant that any pre-existing racial disparities in the criminal justice system were perpetuated and reinforced by the algorithm.

The Wisconsin Supreme Court ultimately upheld Loomis' sentence, ruling that while COMPAS could be used as an advisory tool, it should not serve as the sole determinant in sentencing decisions. The court acknowledged concerns regarding algorithmic opacity but concluded that the use of risk assessments did not, in itself, violate due process, provided that judges exercised independent reasoning. However, the decision left unresolved critical questions about the accountability of algorithmic systems in judicial decision-making and the extent to which they align with constitutional protections.

Beyond legal considerations, the Loomis case underscores the broader ethical dilemma of algorithmic bias in the justice system. The misclassification of defendants based on algorithmic risk scores not only affects individual sentencing outcomes but also undermines public trust in judicial fairness. If algorithmic tools systematically disadvantage certain demographic groups, they risk entrenching historical injustices and eroding the legitimacy of legal institutions. These concerns highlight the urgent need for regulatory oversight, algorithmic transparency, and safeguards to ensure that AI-driven legal tools uphold fundamental principles of justice and human dignity.

The most recent 2024 case of Mary Louis v. SafeRent Solutions exemplifies the profound implications of algorithmic bias in the housing sector, disproportionately

disadvantaging marginalized communities and undermining fundamental principles of human dignity. Mary Louis, a Black woman with a 16-year flawless rental history, was denied an apartment due to an algorithmic assessment conducted by SafeRent Solutions. The algorithm, which heavily weighted credit scores while disregarding housing vouchers, systematically disadvantaged minority applicants, effectively excluding them from equal access to housing opportunities. This case is emblematic of how algorithmic decision-making, when detached from human oversight and contextual nuance, can entrench discriminatory outcomes. By reducing individuals to decontextualized numerical profiles, such automated systems strip away the recognition of personal circumstances, eliminating the possibility of individualized consideration that human dignity demands. As Louis aptly remarked, "Everything is based on numbers. You don't get the individual empathy from them" (Sherman 2024). This absence of human judgment in critical life decisions raises fundamental ethical concerns, as it diminishes individuals to mere data points, disregarding their inherent worth and unique lived experiences.

From a legal standpoint, the case raises pressing issues under the Fair Housing Act (FHA), which prohibits discrimination based on race, national origin, and other protected characteristics. SafeRent's failure to incorporate alternative financial indicators – such as rental history or non-traditional credit metrics – effectively excluded a disproportionate number of minority applicants, challenging the legal principles of fairness and non-discrimination. Moreover, the opacity of the algorithm, coupled with its failure to account for socio-economic realities, exemplifies a broader pattern in which automated decision-making systems operate as black boxes, obscuring the mechanisms through which exclusionary practices occur. This lack of transparency is particularly troubling, as it obstructs individuals' ability to challenge discriminatory outcomes, thereby weakening procedural fairness and due process protections.

Beyond legal considerations, the SafeRent case underscores the societal and dignitary consequences of algorithmic bias in housing. Housing accessibility is not merely an economic issue; it is foundational to human dignity, as it determines one's ability to secure stability, privacy, and social belonging. When automated systems systematically deny minority groups access to rental markets, they exacerbate patterns of segregation and economic disparity, effectively relegating vulnerable populations to a state of algorithmic exclusion. The failure to recognize the human dimension in algorithmic decision-making thus constitutes a direct affront to human dignity, as it deprives individuals of the agency to contest decisions that profoundly affect their lives.

The broader implications of this case demand greater algorithmic accountability and enhanced regulatory scrutiny to prevent the perpetuation of digital discrimination. Fairness-aware machine learning models must be integrated into housing

algorithms to ensure that decision-making processes uphold principles of equality and non-discrimination. Additionally, mechanisms for meaningful human oversight must be embedded into algorithmic governance frameworks, ensuring that technology serves, not undermines, human dignity. The SafeRent case is a stark reminder that algorithmic fairness is not merely a technical challenge but a moral and legal imperative, necessitating a recalibration of AI governance to centre human dignity as a guiding principle in automated decision-making.

The case studies analysed in this section – Meta's targeted advertising, COMPAS in judicial sentencing, and SafeRent's housing discrimination – underscore the far-reaching consequences of algorithmic bias and its profound implications for human dignity. While algorithmic systems promise greater efficiency, their reliance on historical data, opaque decision-making processes, and the absence of meaningful human oversight risk perpetuating systemic inequalities. These biases not only undermine equal access to opportunities but also erode individual autonomy and procedural fairness, core components of human dignity.

Addressing these challenges requires a multi-pronged approach that prioritizes regulatory reforms, algorithmic transparency, and the proactive integration of human dignity as a foundational design principle. Regulatory frameworks must mandate explainability and accountability to prevent automated systems from rendering high-stakes decisions without recourse for affected individuals (Cheng and Liu 2023). Additionally, technical interventions, such as fairness-aware machine learning models and human-in-the-loop oversight mechanisms, must ensure that AI systems recognize the intrinsic worth and rights of individuals, rather than reducing them to decontextualized data points. As algorithmic systems become increasingly pervasive, embedding human dignity into algorithmic governance is no longer optional but imperative. Algorithms must be designed to serve human values rather than subvert them, reinforcing the principles of fairness, justice, and equal treatment. The alignment of AI systems with human dignity is not merely a technical aspiration but a legal and ethical necessity, essential for fostering public trust and ensuring that algorithmic governance operates within the bounds of fundamental rights and social equity.

5 AI Governance Based on Human Dignity

5.1 Enhancing Technological Transparency and Accountability Mechanisms

Ensuring technological transparency and accountability in AI governance is fundamental to safeguarding human dignity. Algorithmic decision-making often

relies on vast datasets, yet the opacity of data and the black-box nature of algorithms can entrench biases and exacerbate social inequalities. Thus, algorithmic governance requires not only technical improvements but also robust institutional safeguards to ensure that AI systems respect individual rights, promote fairness, and prevent discriminatory or unjust outcomes.

One of the most critical measures is strengthening ex-ante controls by democratizing data collection and establishing an effective opt-out mechanism. Since data serves as the foundation of algorithmic decision-making, its collection inherently involves value-laden choices. To mitigate the risks of biased or unfair datasets, AI service providers must adhere to transparent and participatory data governance frameworks before training and deploying algorithms. In particular, service providers should be required to obtain explicit, informed consent from data subjects when collecting personal data, ensuring that individuals understand how their data will be used. Beyond obtaining consent, it is essential to implement robust opt-out mechanisms that empower individuals to control their data. First, data subjects should have the right to selectively provide information without facing service restrictions or discrimination for withholding personal data. Second, individuals must retain the right to request data deletion or restrict the use of their information beyond the original intended purpose, reinforcing personal autonomy over data and preventing its misuse (MacCarthy 2018).

In addition to ex-ante controls, regular independent algorithmic audits should be institutionalized to enhance transparency, fairness, and accountability. Since AI models continuously learn and adapt, new biases may emerge over time, making one-time compliance checks insufficient to address long-term risks. To counteract this, regulatory frameworks should require periodic audits conducted by independent third-party entities, focusing on assessing bias mitigation, transparency, and the broader ethical impact of AI systems. This is particularly crucial for high-risk applications, such as criminal justice, employment, and housing, where algorithmic decisions have profound social consequences. Regulators could mandate periodic fairness reports and bias stress tests to evaluate how AI models perform in different contexts and prevent the inadvertent reinforcement of structural inequalities. Furthermore, it is imperative to establish clear accountability mechanisms that provide individuals with legal recourse if they are negatively impacted by algorithmic bias or discrimination, ensuring that those affected can seek redress and challenge unjust outcomes.

Beyond auditing and accountability, adopting open-source governance models can significantly enhance the transparency and security of AI systems by fostering global collaboration. Open-source governance plays a pivotal role in reducing AI opacity, increasing scrutiny, and enhancing fairness. This approach rests on three key pillars. First, leveraging the open-source community to improve AI safety and fairness. Open-source AI models such as Llama2, Falcon, and Vicuna have

demonstrated how greater transparency enables developers, researchers, and civil society to collaboratively validate models, identify biases, and implement safeguards to ensure that AI development aligns with the public interest. Second, promoting decentralization to counter AI monopolization. Open-source governance can prevent a small number of powerful technology firms from exerting unchecked control over AI development and deployment. This aligns with European AI governance strategies, which emphasize the need for reliable, predictable, explainable, and secure AI models subject to democratic oversight rather than unilateral corporate decision-making. Third, fostering international cooperation to establish global AI governance standards. China's AI open-source governance framework remains in its early stages, with open-source communities, infrastructure, and policy frameworks still developing. To strengthen AI governance on a global scale, China could collaborate with international partners to co-develop open-source governance standards and advance AI alignment frameworks based on measurable, testable, and accountable public goods. Such efforts would ensure that AI governance remains inclusive, transparent, and attuned to ethical considerations.

Lastly, AI governance must incorporate broader public participation and deliberation mechanisms to uphold human dignity. AI regulation should not be dictated solely by technical experts and policymakers; rather, it should involve a diverse range of stakeholders, including developers, academics, civil society organizations, and affected communities. Establishing public forums, open policy consultation mechanisms, and AI ethics committees can create democratic spaces for inclusive debate and collective decision-making. These structures enable the public to exercise their right to know, scrutinize AI decision-making, and demand meaningful explanations for algorithmic outcomes, ensuring that AI systems align with fundamental ethical and human rights principles.

In short, transparency and accountability mechanisms form the foundation of responsible AI governance. Their implementation is not merely a matter of technical optimization but an essential legal, ethical, and societal imperative. By democratizing data governance, institutionalizing independent algorithm audits, promoting open-source governance, and fostering public engagement, AI systems can be made more interpretable, equitable, and aligned with human dignity. These measures are essential in preventing algorithmic bias from perpetuating systemic injustices and ensuring that algorithmic decision-making serves rather than undermines societal well-being.

5.2 Human-oriented AI Design and Application

In the process of AI system design and application, we should adhere to the human-oriented approach to ensure that the development of technology is in line with the

principle of human dignity, taking into account efficiency, fairness and justice, so as to build a sustainable and responsible AI governance system.

Firstly, human dignity values should be embedded in the early stage of AI system design to ensure that algorithms not only pursue optimization and efficiency, but also fully consider individual differences and human needs. This means that the development of algorithms should not only focus on data-driven decision-making capabilities, but should also introduce an ethical review mechanism to avoid unfair impacts of the technology on specific groups. For example, in the selection and processing of training data, the solidification of historical bias should be prevented, while the value considerations of diversity and inclusiveness should be introduced to guarantee the fairness and reasonableness of the system output.

Secondly, human-machine collaborative decision-making mechanisms should be strengthened to ensure that AI operates efficiently while retaining the right to human intervention to ensure that key decisions are always supervised by humans. Fully automated decision-making may lead to problems of unclear responsibility and difficulty in pursuing accountability, so minimum standards for human intervention should be set in areas involving significant social impacts, such as healthcare, justice, and public services. For example, in high-risk decision-making, a "human review" model can be adopted, whereby machines provide advice, but the final decision-making power remains with humans, in order to reduce the risk of algorithmic bias and misjudgement.

In addition, technology should be utilized to promote the construction of "global public goods", i.e., to develop AI governance tools and platforms with universal value, transparency and fairness, so as to serve the common interests of global users. The release of AI governance test frameworks, software toolkits, and other means can help countries test and evaluate AI systems to ensure that they comply with ethical standards, human values and digital human rights indicators. These tools can not only improve the interpretability and transparency of large models, but also establish standard procedures for red team testing, strengthen the value intervention and automated screening of training data, to promote the development of AI in line with the value of human dignity.

Finally, a multi-dimensional evaluation system should be constructed to ensure the legality, ethics and technical feasibility of AI systems. From the legal level, clear compliance standards should be established to ensure that the design and application of algorithms comply with relevant regulations, such as data protection laws and anti-discrimination laws. From the ethical level, an ethical review mechanism should be adopted to ensure that AI systems do not jeopardize individual rights or exacerbate social injustice. From the technological level, the research and development of technologies for interpretability and transparency should be promoted to ensure that the decision-making process of AI is traceable and understandable, and that it can be intervened and corrected when necessary.

5.3 International Cooperation and Intergovernmental Dialogue

The global governance of AI requires countries to strengthen the formulation and implementation of transnational regulatory standards to ensure that human dignity is effectively maintained on a global scale. In the current context of intensified international competition, although there is a game of interests among countries in the formulation of AI rules, international cooperation is still an indispensable key path.

Countries should utilize the intergovernmental dialogue mechanism to promote cooperation among big countries in the field of AI governance and develop a unified cross-cultural and cross-jurisdictional regulatory framework. For example, in terms of data privacy protection, algorithmic transparency, and AI ethics, countries can promote the achievement of global standards through the United Nations, the G20, the OECD, and other international organizations, to ensure that AI research, development, and application will not cause systematic damage to individual rights and social equity. In addition, informal discussions around the banning of autonomous weapons can be initiated among countries mastering cutting-edge AI technology to promote deliberation and exchange on the legal, design, engineering and strategic issues of autonomous weapons, to avoid the risk of AI getting out of control in the military sphere.

Besides, countries should be encouraged to share experiences and data in the field of AI governance and form a global governance alliance to jointly address algorithmic bias and its challenges to social justice and human dignity. For instance, a transnational AI ethics review committee should be established to promote the exchange of experiences among countries in AI compliance review, risk assessment and safety and security to ensure that the development of AI technology is in line with universally accepted ethical and legal frameworks.

Furthermore, in terms of realistic paths, China should play a leading role in digital human rights issues and push developing countries to reach a minimum consensus. For example, in terms of the safety and stability of AI technology, value alignment, etc., China can establish a regional AI digital human rights governance network with the EU and APEC to promote the synergistic development of technical standards. At the same time, China can also advocate the “right to access advanced AI” and the “right to share AI dividends” and promote transparency and openness of developed countries’ AI platforms, technologies and information, to reduce the AI technological divide and promote fair development on a global scale. This will reduce the AI technological divide and promote equitable development on a global scale.

5.4 Public Participation, Education, and Training

AI governance is not only about government regulation and technological innovation, but also requires extensive public participation and cognitive enhancement. Therefore, the public's right to information and supervision should be strengthened to ensure that all sectors of society can effectively participate in the AI decision-making process, to enhance the overall awareness of the risk of algorithmic bias and its impact on human dignity.

A transparent AI decision-making disclosure mechanism should be established to enable the public to understand the operation, decision-making logic and potential risks of AI systems. For example, platforms can regularly publish AI transparency reports that disclose the results of fairness tests of algorithms, possible bias problems, and improvement measures. Besides, public feedback channels can be established to enable users to comment on the operation of AI systems and apply for review or correction of erroneous decisions when necessary.

Moreover, interdisciplinary talent training should be promoted in the fields of law, ethics, and computer science to improve practitioners' and users' awareness of and ability to respond to the risk of algorithmic bias. For technology practitioners, ethical training should be strengthened in the process of AI development and application so that they can incorporate fairness and transparency considerations in the algorithm design phase. For ordinary users, basic knowledge of AI should be popularized through educational institutions, online courses and public forums to enable the public to better understand and monitor AI systems.

In addition, value alignment mechanisms in human-computer collaboration should be promoted to ensure that AI systems truly reflect human values and needs. Best practices include building computational frameworks for bidirectional value alignment, facilitating real-time human-computer communication, and enhancing AI's ability to understand human intentions. For example, algorithm optimization methods based on the RICE principle can be used to ensure that AI can dynamically adjust to different ethical and cultural contexts during the decision-making process. At the same time, in the process of human-machine collaboration, it must be ensured that the final decision-making power remains in human hands to prevent the autonomy of AI from leading to uncontrollable ethical risks.

In conclusion, AI governance requires a multidimensional approach that integrates systemic, technological, and societal efforts to ensure the controllability, transparency, and fairness of AI systems. By embedding human dignity into AI design, fostering international cooperation and intergovernmental dialogue, and enhancing public participation through education and training, we can facilitate

meaningful value alignment in AI development. This holistic approach not only mitigates the risks of algorithmic bias but also safeguards social equity and individual rights, ensuring that AI serves as a force for inclusive and ethical progress.

6 Conclusions

In the era of global competition in the digital economy, compared with technology and capital, the construction of governance rules has become a new power discourse in the digital age, playing a substantial role in core strategic functions. In the context of AI governance, algorithm governance is an issue involving complex interactions among multiple subjects such as individuals, enterprises, and public institutions. Under this realistic challenge, it should be realized that the effective implementation of the right to interpret algorithms requires an appropriate supporting system as a basis, and human dignity, as the core of legal regulation, can provide legislators with an internal and external perspective beyond the idea of rights construction, and achieve a synergistic mechanism of organic combination of law and technology through value alignment to maximize the effectiveness of the system. EU law, represented by GDPR and case law, has not only worked hard in the field of personal privacy protection, but also rapidly iterated in the regulation of automated decision-making, revealing the demand orientation behind AI that is aligned with the value of human dignity. In other countries, a series of laws related to data security and personal information protection have been gradually introduced in recent years, demonstrating the determination and wisdom of legislators in major countries in the world in the field of digital technology governance.

It can be foreseen that the data and algorithm governance systems of more and more countries will be improved day by day. The information technology society is often described as the "post-factual era" of legislation, and each technology is increasingly showing a butterfly effect that affects the entire body. Overcoming the passive pursuit and delayed response to technology in legislation and properly balancing the interests of multiple subjects depends on the improvement of legislative technical capabilities, as well as the active and collaborative participation of industry organizations, enterprises and individuals. Among all governance tools, grasping human dignity as the principle of the value of AI, responding to algorithmic risks and being free from the constraints of automated decision-making, and improving the internal structure of rights and the institutional environment for external implementation are of great and far-reaching significance for moderately regulating and preventing the technical risks of algorithmic discrimination and countering the digital survival dilemma of algorithmic manipulation.

Research ethics: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Conflict of interest: The authors state no conflict of interest.
Research funding: This work was supported by the project of National Social Science Foundation (Grant No. 24BYY151) and the Fundamental Research Funds for the Central Universities, Zhejiang University.
Data availability: Not applicable.

References

Bean, E., C. Burleigh, C. Haskell, T. Burris-Melville, J. Payne, and B. Pathak. 2025. "Eavesdropping on UNESCO AI Policy, Leadership, and Ethics." *Journal of Leadership Studies* 18 (4): 98–110.

Bioy, X. 2006. "L'identité de la Personne Devant le Conseil Constitutionnel." *Revue Françai se De Droit Constitutionnel* 65 (1): 73–95.

Botha, H. 2009. "Human Dignity in Comparative Perspective." *Stellenbosch Law Review* 20 (1): 169–95.

Brennan, T., W. Dieterich, and B. Ehret. 2009. "Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System." *Criminal Justice and Behavior* 36 (1): 21–40.

Burgess, J., N. Carah, D. Angus, A. Obeid, and M. Andrejevic. 2024. "Why Am I Seeing This Ad? The Affordances and Limits of Automated User-Level Explanation in Meta's Advertising System." *New Media & Society* 26 (9): 5130–44.

Cañamares, Rocío, and Pablo Castells. 2018. "Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems." In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. Anna Arbor MI, 415-24. New York: Association for Computing Machinery.

Chang, Y., X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, and H. Chen, et al. 2024. "A Survey on Evaluation of Large Language Models." *ACM Transactions on Intelligent Systems and Technology* 15 (3): 1–45.

Cheng, L., and X. Gong. 2024. "Appraising Regulatory Framework Towards Artificial General Intelligence (AGI) Under Digital Humanism." *International Journal of Digital Law and Governance* 1 (2): 269–312.

Cheng, L., and X. Liu. 2024. "Unravelling Power of the Unseen: Towards an Interdisciplinary Synthesis of Generative AI Regulation." *International Journal of Digital Law and Governance* 1 (1): 29–51.

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

Dearing, A. 2017. *Justice for Victims of Crime*. Vienna: Springer.

Dong, Hanze, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pab, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. "Raft: Reward Ranked Finetuning for Generative Foundation Model Alignment." *arXiv preprint* arXiv:2304.06767. https://doi.org/10.48550/arXiv.2304.06767.

Engstrom, D. F., and D. E. Ho. 2020. "Algorithmic Accountability in the Administrative State." *Yale Journal on Regulation* 37: 802–54.

Gabriel, I. 2020. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30 (3): 411–37.

Gish, J. J., C. M. Barnes, A. Gupta, and K. Nair. 2023. "Presumed Patriarchy: How a CEO's Masculine Appearance Affects Perceptions of Sexual Harassment in Organizations." *Journal of Management* 51 (2): 812–42.

Goodman, M. D. 2005. "Human Dignity in Supreme Court Constitutional Jurisprudence." *Nebraska Law Review* 84 (3): 743–64.

Günther, M., and A. Kasirzadeh. 2021. "Algorithmic and Human Decision Making: For a Double Standard of Transparency." *AI & Society* 37 (1): 375–81.

Habbal, A., M. K. Ali, and M. A. Abuzaraida. 2024. "Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, Applications, Challenges and Future Research Directions." *Expert Systems with Applications* 240: 122442.

Hernandez, J. G. 2015. "Human Value, Dignity, and the Presence of Others." *HEC Forum* 7 (3): 249–63.

Higgins, K. C., and S. Banet-Weiser. 2024. "A Roundtable Discussion of Kathryn Claire's Believability: Sexual Violence." *Media and the Politics of Doubt Feminist Theory* 25 (3): 263–87.

Hoofnagle, C. J., B. van der Sloot, and F. Z. Borgesius. 2019. "The European Union General Data Protection Regulation: What it Is and what it Means." *Information & Communications Technology Law* 28 (1): 65–98.

Larson, J., J. Angwin, S. Mattu, and L. Kirchner. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*, May 23, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed March 22, 2025).

Leal, M. C. H., and D. S. Crestane. 2023. "Algorithmic Discrimination as a Form of Structural Discrimination: Standards of the Inter-American Court of Human Rights Related to Vulnerable Groups and the Challenges to Judicial Review Related to Structural Injunctions." *Unio - EU Law Journal* 9 (1): 29–44.

Lehdonvirta, V., B. Wú, and Z. Hawkins. 2024. "Compute North vs. Compute South: The Uneven Possibilities of Compute-Based AI Governance Around the Globe." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (1): 828–38.

Li, S., Ramakrishnan, N. 2025. "Oreo: A Plug-in Context Reconstructor to Enhance Retrieval-Augmented Generation." *arXiv preprint* arXiv:2502.13019.

Llano, F. H. 2019. "Transhumanism, Vulnerability and Human Dignity." *Deusto Journal of Human Rights* 4: 39–58.

Longo, E. 2022. "The Risks of Social Media Platforms for Democracy: A Call for a New Regulation." In *Law and Artificial Intelligence. Information Technology and Law Series*, edited by B. Custers, and E. Fosch-Villaronga, 169–86. The Hague: TMC Asser Press.

MacCarthy, M. 2018. "Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms." *Cumberland Law Review* 48 (1): 67–148.

Maiti, M., P. Kayal, and A. Vujko. 2025. "A Study on Ethical Implications of Artificial Intelligence Adoption in Business: Challenges and Best Practices." *Future Business Journal* 11 (34): 1–12.

Malgieri, G. 2019. "Automated Decision-Making in the EU Member States: The Right to Explanation and Other "Suitable Safeguards in the National legislations"." *Computer Law & Security Report* 35 (5): 105327.

Mary Louis v. SafeRent Solutions. 2024. Case No. 1:20-cv-03142 (D. Colo. Nov. 15, 2024). https://www.courtlistener.com/docket/18162917/louis-v-saferent-solutions-llc/ (accessed March 22, 2025).

Mayson, S. G. 2019. "Bias in, Bias Out." *Yale Law Journal* 128 (8): 2218–300.

Molina, A. U., and M. H. Subias. 2024. "Personalization of Content in Video-on-Demand Services: Insights from Satisfaction over Social Media Algorithms." *Revista ComHumanitas* 15 (2): 175–87.

Nardone, V. 2019. "The Passenger Name Record Case: Profiling Privacy and Data Protection Issues in Light of CJEU's Opinion 1/15." In *Use and Misuse of New Technologies*, edited by E. Carpanelli, and N. Lazzerini, 135–50. Cham: Springer.

Neuwirth, R. J. 2023. “Equality in View of Political Correctness, Cancel Culture and Other Oxymora.” *International Journal of Legal Discourse* 8 (1): 1–29.

Ng, G. W., and W. C. Leung. 2020. “Strong Artificial Intelligence and Consciousness.” *Journal of Artificial Intelligence and Consciousness* 7 (1): 63–72.

Ord, T. 2020. *The Precipice: Existential Risk and the Future of Humanity*. New York: Hachette.

Proshansky, H. M. 1978. “The City and Self-Identity.” *Environment and Behavior* 10 (2): 147–69.

Rao, N. 2012. “American Dignity and Healthcare Reform.” *Harvard Journal of Law and Public Policy* 35 (1): 173–85.

Rjoub, G., J. Bentahar, O. Abdel Wahab, R. Mizouni, A. Song, R. Cohen, H. Otrok, and A. Mourad. 2023. “A Survey on Explainable Artificial Intelligence for Cybersecurity.” *IEEE Transactions on Network and Service Management* 20 (4): 5115–40.

Romero Moreno, F. 2024. “Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content.” *International Review of Law Computers & Technology* 38 (3): 297–326.

Sengupta, S. 2025. “Beware of First-Hand Ideas: Foretelling the Existential Risk of Technology Deification and Dependence through Forster’s Narrative Lens.” *AI & Society*, https://doi.org/10.1007/s00146-025-02228-7.

Shah, K., H. Joshi, and H. Joshi. 2025. “Integrating Moral Values in AI: Addressing Ethical Challenges for Fair and Responsible Technology.” *Journal of Informatics and Web Engineering* 4 (1): 213–27.

Sherman, N. 2024. Judge Approves Settlement in AI Discrimination Lawsuit over Rental Scoring Algorithm. *Lawyer Monthly*. https://www.lawyer-monthly.com/2024/11/judge-approves-settlement-in-ai-discrimination-lawsuit-over-rental-scoring-algorithm/ (accessed March 20, 2025).

Simmons, R. 2018. “Big Data and Procedural Justice: Legitimizing Algorithms in the Criminal Justice System.” *Ohio State Journal of Criminal Law* 15 (2): 573–81. https://hdl.handle.net/1811/86574.

Sparks, J., and A. T. Wright. 2025. “Models of Rational Agency in Human-Centered AI: The Realist and Constructivist Alternatives.” *AI Ethics*. https://doi.org/10.1007/s43681-025-00658-z.

State v. Loomis. 2016. *881 N.W.2d 74*. https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690 (accessed March 21, 2025).

Tie, Guiyao, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, et al. 2025. “A Survey on Post-training of Large Language Models.” *arXiv preprint* arXiv:2503.06072. https://doi.org/10.48550/arXiv.2503.06072.

U.S. Department of Justice. 2022. Justice Department Secures Groundbreaking Settlement Agreement with Meta Platforms, Formerly Known as Facebook, to Resolve Allegations of Discriminatory Advertising. https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known (accessed March 20, 2025).

Veale, M., and Z. F. Borgesius. 2021. “Demystifying the Draft EU Artificial Intelligence Act – Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach.” *Computer Law Review International* 22 (4): 97–112.

Wu, Sihao, Xiaonan Si, Chi Xing, Jianhong Wang, Gaojie Jin, Guangliang Cheng, Lijun Zhang, and Xiaowei Huang. 2025. “Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing.” *arXiv preprint* arXiv:2502.07829. https://doi.org/10.48550/arXiv.2502.07829.

Zerilli, J., A. Knott, J. Maclaurin, and C. Gavaghan. 2018. “Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?” *Philosophy & Technology* 32 (4): 661–83.

Bionotes

Yilin Zhao
Guanghua Law School, Zhejiang University, Hangzhou, China
21905065@zju.edu.cn
https://orcid.org/0000-0002-7927-2660

Yilin Zhao is Research Fellow in the Guanghua Law School, Zhejiang University. Her research fields include international law, global governance, and cyber laws.

Zeshen Ren
Faculty of Law, National University of Singapore, Singapore, Singapore
e1291370@u.nus.edu
https://orcid.org/0009-0004-8626-3705

Zeshen Ren is a PhD student at the Faculty of Law, National University of Singapore. Her research focuses on the legal and governance challenges of emerging technologies.