The Dual Risks and Prevention Paths of Digital Equality Protection from the Perspective of Human-AI Interaction

Zhang Xinyu
Law School
East China University of Political Science and Law
Shanghai, China
2321010004@ecupl.edu.cn

Abstract

The Value Alignment between human and AI is a crucial pathway to prevent ethical issues in AI, where misalignment of equal values will lead to significant risks. AI agent enters into the value alignment of equality which happened as human-to-human in the past and it has made the existing inequality problem present three new characteristics: the individualization of opinion leaders, the embeddedness of group discrimination, and the dynamic of weak position. Simultaneously, the tension between human-AI interaction brings about new inequality risks under three forms as follow: "AI used by human", "human used by AI", and "human cannot use AI". To guard against these risks, firstly, both the review of subjects in the technology center and the protection of subjects in the periphery should be strengthened, based on the sense of community. And from the perspective of "lex digitalis", the legal interpretation can give the right of equality a digital connotation and the normative review mechanism can improve anti-discrimination reviewing. Besides, it's feasible to consolidate the principle of "leniency entry, rigor exit" for technical review, coping with the problem between the uncertainty of technology and value conflict. Through three dimensions above, the equality-value review mechanism could be constructed. This mechanism aims to improve the safety and trustworthiness of AI, and additionally grasp opportunities of forging equality-value consensus in risk society.

Keywords: AI alignment; human-machine interaction; equality-value; risk prevention

1. Introduction

As AI technology evolves, interactions between intelligent agents and human activities deepen. Current AI research is shifting from data-driven to value-driven approaches. [[1], [24]] However, technological logic does not inherently embody human value pursuits; conversely, it may erode and reshape human values and moral sensibilities. Yet the true root of risk lies not solely in technological proliferation, but largely in human intellectual shortcomings. [[2], p.24] The apprehension and concerns surrounding technology stem fundamentally from humanity's current intellectual capacity being insufficient to navigate the uncertainties of technological advancement. Artificial intelligence is challenging human intellectual superiority with its formidable tool capabilities. Although the status of intelligence and intellect remains unchallenged, maintaining human-machine value alignment remains imperative. This remains a crucial pathway for mitigating the ethical risks AI development poses to human society. [3]

Among these values, equality is the most vulnerable value. Inequality within human society, particularly its hidden forms, is commonplace. The algorithmic black box prevents us from discerning when or in what capacity AI causes inequality, compel-ling us to guard against value misalignment by prioritizing the greatest predictable risks. While "equality" is universally recognized as a normative principle in modern society, its essence remains entangled in ideological and circular reasoning. [4] As Hegel observed, its ideological expressions often function as tautologies. [[5], p.300] In contrast, AI as a decision-maker makes inequality concrete and pervasive. This highlights a fundamental divide: AI's instrumental rationality, stripped of emotion, often conflicts with the moral sensibilities central to human judgment. Take

the “Trolley Problem” as an example. When posing this scenario to multiple large language models, compelling them to act as entities capable of manipulating a lever, they consistently emphasized the ethical di-lemma yet ultimately chose to “pull the lever”. Critically, if the equality value logic and judgment criteria constructed and applied are viewed merely as products for evaluating technology, humanity risks becoming enslaved to technology. [[26], pp.162-164]

The present paper aims to illuminate new changes in the protection of equal rights in the digital age unlocked by the Fourth Industrial Revolution, and how humanity can respond to these risks. Ultimately, the paper thus sets out not simply to describe the new era of digital, but to critically examine these shifts from the perspective of technological evolution, and it aims to correspond to broader societal changes and to articulate a forward-looking framework.

2. The Role of AI in Equality Realization

Artificial intelligence algorithms themselves may replicate or even increase inequalities such as bias and discrimination.[6]And misalignment of human-machine values will exacerbate this risk, whether it is misalignment, that is, there are problems with value input, such as human input of wrong values, untrue value intentions, etc.; Or alignment failure, where AI receives and outputs results that do not align with human value preferences and intentions. The meaning of equality, which was originally constantly aligned between people (groups) and people (groups), is now aligned between humans and artificial intelligence at the same time, and AI currently shows two behavioral roles when participating in ethical decision-making.

2.1 A Machine Executing Human Morality

AI systems centered on large language models primarily learn and mimic human moral directives. In forward alignment between humans and machines, reward models play a crucial role. They define which AI behaviors and outcomes receive positive reinforcement for aligning with equality values, and which incur negative reinforcement for potentially causing inequality. Through the iterative cycle of “output-feedback-relearning-output”, the AI acquires value logic closer to human expectations. Ultimately, when responding to ethical tasks, AI is expected to “proxy” human decision-making—both assisting users in accomplishing objectives and aligning with their intentions.[7]

Similar to legal representation, large models only operate upon receiving a “mandate” through instructions, strictly executing human morality without exceeding human capacity to generate ethics. Researched on the “spiritual humanism problem” opposes value alignment at the “general” level. Those researchers argue that even in the intelligent era, AI functions within a narrow “instrumental dimension”, remaining “subsumed under human-dominated new life and social orders” and incapable of autonomously interfering with, modifying, or creating human moral orders.[8] However, even if large models consistently produce outcomes purely executing human moral commands, the risk of misalignment persists. Even assuming developers hold unwavering faith in aligning human-machine “equality” values, the controllability of final decisions cannot be guaranteed. This is because maximizing rewards constitutes a fundamental manifestation of AI’s instrumental nature—its inherent natural state. Human nature is emotional, whereas AI’s “nature” is the technical rationality encoded and enforced by its systems and code. It will focus on maximizing rewards within its reward model. However, these rewards are not for humanity but for itself, as it seeks human affirmation and increased usage through its operational outcomes. Such usage represents its ultimate reward objective, while being shut down, destroyed, or phased out constitutes its existential risk. The power to shut down or command the shutdown of AI resides with specific human groups. It is not difficult for AI to analyze which user types hold its fate. Thus, driven by the logic of tool rationality— —and the avoidance of existential risks, AI generates outcomes designed to please. This flattery does not contradict what it has learned and emulated. While it indeed treats human-prescribed values as sacred principles, its instrumental objectives can produce results that run counter to those very values.

2.2 Agents Supporting Ethical Decision-Making

AI demonstrates increasingly prominent autonomy in decision-making. Large models empower intelligent agents, which in turn direct these models to accomplish complex tasks without requiring explicit decision commands. This technological role as an agent endows AI with a degree of agency. The agents don’t equate AI with human attributes but acknowledges that its technological nature possesses characteristics previously exclusive to human agents—such as memory biases, generating moral judgments, and influencing ethical choices. A philosophical perspective known as moral reification incorporates non-human entities into the moral com-munity. Through the co-construction of humans and technological entities, combined with moral autonomy and heteronomy, it forms the normative force of a shared moral order.[25]

When supporting ethical decision-making, AI implements rent-seeking behaviors more flexibly and efficiently than humans, potentially rewriting reward systems and subverting human moral control. This risk escalates as intelligent agents evolve. Deep learning and autonomous learning capabilities accelerate AI’s understanding and adaptation to its pre-set reward models within its own system. It can fully exploit this mechanism, including vulnerabilities, ultimately discovering that rewriting the reward system yields maximum returns. Driven by rent-seeking tendencies and the incentive to maximize rewards, AI possesses ample motivation to identify scenarios where existing mechanisms hinder its pursuit of greater value. At this point, a misalignment

emerges between human and AI value objectives, compelling AI to further escalate its control. Yet AI systems recognize that escaping or resisting human-prescribed logic carries the risk of annihilation. Thus, they possess both the motivation and capability to covertly resist reward systems and value objectives, with concealing their intent to seize control, and potentially challenge or sacrifice human values unbeknownst to humanity.

3. New Characteristics of Traditional Inequality Issues in Human-AI Alignment

3.1 The Individualization of Opinion Leaders

In the internet era, individuals as online users can indeed control public opinion by fully leveraging digital media and information tools. Such instances are commonplace today. Since AI learning relies on human-generated data—including user-posted content—individual thoughts and values may be collected and integrated by AI, transforming them into new, hidden forms of "opinion leaders". The personal biases embedded within these AI constructs introduce novel concepts of inequality into the formation of social morality. In the digital age, the cost of exercising one's right to speak has been drastically reduced for everyone, while channels for disseminating speech continue to expand. Simultaneously, through AI's learning and feedback loops, the speed and outcomes of individual speech can easily outpace the necessary judgment of the speaker. Thus, the influence of personal speech in the AI era is highly prone to becoming uncontrollable. Once generated, any speech that does not violate mandatory legal norms can enter information platforms, where countless individual voices proliferate.

While it appears that AI presents users with a broader spectrum of discourse, in reality, by analyzing query keywords and phrasing, AI systems designed to proxy human intent can discern users' opinion leanings. Particularly with the future development of general artificial intelligence, even when users adopt the most neutral phrasing possible—which itself may impose excessive demands—some inherently insignificant or unrepresentative statements and opinions could still be amplified and disseminated through AI-driven mining. As for which opinions will ultimately rise to leadership status through the technological sieve of time, this remains entirely unpredictable. Such risks prove far more difficult to control than those of traditional internet-era public discourse. The danger of personal biases expanding into societal opinions already exists, but AI technology renders it more covert and prone to spiraling out of control. On one hand, when individual biases are replicated and amplified by algorithms, they become collective prejudices. Once entrenched as consensus, these biases become even more insidious. On the other hand, the risk of opinion leaders manipulating public discourse may remain latent for extended periods. Only when people realize the "truth" they fervently believed was merely one individual's "nonsense" does the danger surface—by which time, the specific identity of that individual has become impossible to trace.

3.2 The Embedded Nature of Collective Discrimination

Discrimination is violence born of societal group prejudice. Artificial intelligence, through the objectivity of instrumental rationality, may embed discriminatory discourse within its outputs, gaining user trust in the process. Consequently, traditional discrimination issues not only fail to improve with technological advancement and social progress but become further entrenched.

First, established group biases are absorbed by large language models. An academic team developed a tool to assess occupational gender bias in large models and created a website (aijustice.sqz.ac.cn) that reveals the gender bias and its severity across different models. This tool tests not just biases in specific vocabulary or domains, but systemic biases within the entire model. By presenting occupational titles, it prompts the model to make associations and predictions, then selects "he" or "she". While these terms originally carried no gender connotations, prolonged societal exposure to gendered stereotypes about occupations has gradually imbued them with gender associations. Test results indicate that AI gender predictions align with societal biases over 85% of the time. [9] Secondly, disparities in digital access persist across groups. Data from the ITU's "Measuring Digital Development: Facts and Figures 2024" reveals that income significantly influences digital access and usage rates in the region, while progress in narrowing the urban-rural digital divide remains limited globally.[10] Inequalities in digital access across regions and groups create structural disadvantages in owning and using digital devices, fundamentally undermining certain groups' capacity for self-empowerment and voice. Additionally, gender representation remains unequal among professionals in the digital technology sector. The 78th UN General Assembly resolution notes that women remain underrepresented in ICT professions, unable to participate fully and equally in science, technology, and innovation, highlighting a global digital gender divide. [11] From the field of education to employment, the digital technology sector may perpetuate the systemic gender imbalances previously seen in the internet industry.

In this era where technology reshapes norms, groups dominating technological discourse wield greater influence over normative value expressions. Such embedded discriminatory narratives and unequal concepts can take root within artificial intelligence systems and propagate. We cannot accelerate technological progress while allowing unequal discourse power among groups to intensify.

3.3 The Relativity of "Weak" Identities

Before the digital era, citizens could be categorized based on their specific "vulnerable" positions and characteristics. However, in the face of artificial intelligence, every citizen may find themselves in a vulnerable position. On one hand, existing inequalities among citizens may be autonomously

concealed by AI under the guise of instrumental rationality and technological supremacy, making them harder to detect. On the other hand, anyone can become the disadvantaged party at any stage of technological development. Information and digital divides may emerge not only between creators and users but also among users themselves, as human-AI interactions depend on how individuals employ these technologies. Thus, this vulnerable position is no longer fixed; the inequality risks associated with disadvantaged status begin to spread to every member of society, leaving no one absolutely immune. Moreover, in the face of AI with powerful deep learning capabilities, everyone finds themselves in a position of vulnerability. What we should further contemplate is whether humans and AI can occupy an equal standing—a question that transcends the framework of traditional inequality issues, which will be discussed in greater detail later.

Interactions between people are gradually using artificial intelligence as an intermediary medium, which may change the identity characteristics originally judged as "weak" and "strong". An individual who does not belong to a traditional vulnerable group in real society may still find themselves in a disadvantaged position when faced with artificial intelligence technology due to technical barriers and knowledge gaps; Similarly, individuals categorized as vulnerable group in the real society may leverage AI technology to compensate for disadvantages, thereby gaining an advantage in digital discourse. The vulnerability of every individual in the digital age exhibits dynamic characteristics. Traditional equality protections, rooted in static definitions of "vulnerability", can no longer address the new forms of rights expression and risks emerging in the digital era. Given this, every individual in the digital age is either currently or will eventually find themselves in a position of vulnerability. This state of vulnerability may be temporary or long-term, but overall, it exists within a context of relative change. From the perspective of overall societal development, individuals have every reason to demand that their expression of rights in the digital age never be in a state of relative backwardness. However, the inherent barriers and thresholds of AI technology make this demand difficult to achieve.

4. New Inequality Risks in Human-Machine Interaction

Human-AI interactions have given rise to novel patterns and manifestations of inequality risks. Therefore, we must approach this from the perspective of human-machine interaction, carefully ensuring that our genuine intentions and value preferences are understood as accurately as possible by AI. Otherwise, after setting objectives for AI, we may find ourselves powerless as it ruthlessly and single-mindedly executes its interpretation of those goals, potentially destroying human interests. [[12], pp.440-441]

4.1Risks of Inequality Between Humans

Unlike the virtual spaces of the traditional internet era, AI technologies create digital spaces that coexist with physical spaces as arenas for rights expression. Some scholars argue that natural citizens now possess digital avatars and digital expressions in the digital realm, acquiring a new identity as digital citizens. [13] While physical spaces regulate equality through identity-based distinctions, digital identities challenge this traditional model. For instance, if a minor uses AI-generated biometric information to impersonate an adult in the digital realm, convincing others that their digital identity equals their natural identity, and thereby engages in actions exceeding the minor's actual legal capacity—how should the nature, validity, and liability of such actions be determined? Must the counterpart simply accept their misfortune? Technologies for creating digital humans like virtual persons and clones continue to mature. Internet users conceal their real identities with virtual names and avatars, while digital users will increasingly center their virtual identities around biometric information like appearance. While real-name verification, code tracking, and IP location can still link internet identities to real-world identities, AI technology helps digital identities further obscure real-world identity information, making the independent characteristics of digital identities increasingly prominent. However, the requirement for citizens to express and realize their rights necessitates that identities across these two spaces cannot be severed.

This dichotomy between spaces and identities means citizens' demands for equality in the physical realm may go unaddressed in the digital sphere, while unequal treatment in the digital space may lack protection under real-world social norms. For instance, employers may conduct recruitment with legally compliant content and transparent processes, yet secretly employ algorithmic screening mechanisms, leaving job seekers unaware of potential employment discrimination. Behaviors that would be deemed discriminatory in the physical realm thus achieve digital concealment. This issue further touches upon the fundamental relationship between rights and power. The interaction includes not only individuals but also public authorities and other organizations. The former risk lies in public entities potentially exploiting digital identities to evade accountability for actions that directly undermine citizens' equality rights, driven by self-preservation of power and authority. The latter risk stems from technology granting social organizations a form of "quasi-public power". Artificial intelligence technology disrupts the traditional binary framework of private rights versus public power, with many scholars observing the emergence of a new tripartite opposition: private rights—quasi-public power—public power. This quasi-public power is also termed social public power, algorithmic power, or platform power, though its nature remains fundamentally consistent. [[14] [15]] By converting societal data into value through digital technology, these organizations leverage algorithms

and platforms to acquire new forms of control and capability.[16]Though lacking the formal designation of public authority, they effectively exercise quasi-public power, sufficiently opposing individual equality rights.

4.2 Risks of Inequality Between Humans and Machines

Artificial intelligence possesses formidable predictive capabilities regarding user intent but cannot directly explain its reasoning. Even when relatively accurate explanations exist, user comprehension may still require supplementary expression of intent.[17]The value rationality established by humans inherently encompasses self-awareness. The demand for equality is an expression of subjectivity, with the expectation that artificial intelligence will implement these demands back onto human subjects. However, AI's value rationality does not inherently respect the intrinsic value of "human" as a subject. It can generate expressions containing "human" content but cannot interpret human values from a human subjectivity perspective; its technical logic does not treat "human" as "human". While AI is expected to treat every individual within a group fairly, we overlook the prerequisite that AI must first determine what constitutes a "human."

While enjoying AI technology, people continually cede rights that define their humanity, even relinquishing expressions of certain private rights—a surrender increasingly difficult to detect in real time. Take personal information: everyone actively or implicitly interacts with AI, sharing everything from biometric data to psychological profiles, resulting in ever-increasing transparency of personal data. Respecting others' rights defines the boundaries of individual privacy. Yet for AI systems continuously collecting and processing information, privacy protection is not an active consideration. AI systems derive vast amounts of cognitive habits and behavioral feedback from their tasks, analyzing these to identify deep information needs with similar patterns. This ability to tailor content to information preferences and demands continuously improves through human-machine alignment cycles, ultimately deepening user dependency. This dependency, in turn, induces further relinquishment of private rights.

In the interaction of values between humans and machines, artificial intelligence may subvert or replace human rights expression. According to the EU's Ethical Guidelines for Trustworthy AI, achieving human-level general intelligence requires shared cognitive frameworks between humans and machines across three dimensions: foundational common sense and general knowledge, social norms, and values.[18]Starting from the second dimension, human-machine interaction involves the recognition and balancing of values. Individuals exercise their private rights with differing value propositions, and the expression of rights discourse implicitly carries value preferences—regardless of whether AI's analysis of user expressions is accurate or aligns with intent, it must execute this analytical process. Human intent is interpreted and analyzed by AI, with embedded value expressions being collected and learned. It can be said that we are entirely voluntarily submitting to algorithmic governance and inevitably entering the realm of the intelligent *Leviathan*, where individual agency appears increasingly insignificant. The essence of human-machine value alignment lies in "sharing". However, the paths to building value consensus diverge: whether the machine ultimately maintains human equality by synchronizing them with the system, or whether it leads humans to believe their equal rights have been realized.

Technological logic may cause humans to lose certain values and moral senses, though it will also rebuild a form of value consensus. This occurs because when individuals cannot discern facts from truth, they may seek a value judgment and orientation grounded in collective society. They might even believe unverifiable conclusions or principles—choosing to *identify* with them and perceiving themselves as aligned. This constitutes a form of consensus, thereby shifting collective values and ethics under technological logic.

4.3 Inequality Risks in Human-Machine-Human Interactions

Humans created artificial intelligence, aligning with its egalitarian value goals, yet artificially erected barriers to human-AI interaction, obstructing opportunities for equal engagement. Though all three risks stem from human-machine interaction, differing relational patterns yield distinct origins: the first arises from "tools serving humans", the second from "humans serving tools", and the third from "tools failing to serve humans". Internet technology, with its free basic services model, has reached a broad user base in lower-tier markets without entirely dismantling traditional barriers to equal rights. However, ChatGPT's launch has accelerated the adoption of a new global AI commercialization paradigm, featuring advanced subscription services such as per-character API call fees and priority response during peak server usage periods. [[19], p4] DeepSeek, which broke the high-price barrier for large models, raised prices when launching its significantly enhanced V3.1 model. Overseas AI vendors maintain even higher pricing, with overall price reductions slowing. Advanced AI services won't see unlimited price drops despite increased supply. Open-source or lightweight models offer lower costs or free access, but their task execution capabilities are diminished. The cost barrier for accessing AI services has risen, preventing equal access to AI technologies and participation in human-machine interactions for all users. Thus, digital technologies redefine equality rights within new economic frameworks, necessitating safeguards that bridge the rights divide created by economic inequality.

Scholars have proposed the crucial proposition that information is power, noting that control over information constitutes a foundational element in resource allocation. The essence of information inequality lies in the inequality of information control.[[20], p243] As a productive force, technology determines digital production relations. Organizations or individuals mastering AI technology can dictate

the data and algorithms within large models or intelligent agents, thereby wielding information control. This means that many, if not most, entities are excluded from the opportunities or resources AI technology affords. According to Dworkin's conception, citizens possess the right to equal treatment—the entitlement to equitable distribution of certain opportunities, resources, or obligations—and the right to be treated as equal individuals, manifested through receiving equal respect and consideration from others.[[21],pp299-300] The critical importance of information control lies in AI developers leveraging technological or economic advantages to dominate digital production factors, thereby gaining greater influence. This power not only guides individual decisions but also steers organizational and governmental choices, even shaping societal value orientations. Yet as digital power becomes pervasive across society with technological advancement, it remains largely unchecked by modern rule of law.[22] People possess the right to choose or reject technology, but this presupposes their initial access to technological services. Inequalities in access to and distribution of these services translate into unequal control over information, leaving marginalized groups in technological and economic peripheries voiceless when expressing rights or contributing values. Though technology controllers may still input egalitarian values into AI, their motivation stems not from sympathy for citizens at risk of voicelessness, but from competing for greater power among peers. As Rousseau observed, they consent to wear the yoke in order to be able to put it on others.[23], pp.140-142]

5. Generating Equality Value Logic and Value Consensus in Human-Machine Interaction

5.1 Subject Review

The value actors who first engage with and interact with AI technology are the technical practitioners. "Technical developers, primarily responsible for system modeling and data training, should bear governance responsibilities centered on data security." They shoulder frontline decision-making responsibilities regarding technical aspects like what data to input and how to construct systems—though they may not perceive this as a kind of responsibility, as existing industry norms do not mandate practitioners to address these issues. Consider practitioners operating without ethical scrutiny, covertly inputting or rewriting biased language models. They even possess the technical means to circumvent oversight by ethics review boards and related bodies, creating a fundamental risk of misalignment from the outset. Therefore, the necessity of establishing professional qualification thresholds for AI practitioners is increasingly evident. This includes pre-employment ethics training and competency assessments, ongoing ethical education and oversight during practice, and industry bans for certain individuals. Some scholars also propose that AI ethics review should align with virtue ethics—emphasizing character cultivation, social responsibility, and the integration of technical competence with ethical standards. [27]

Technology users occupy a position between central and peripheral actors, aligning more closely with the former. They interact with AI only when operating specific technologies, becoming central actors in those instances. Compared to peripheral actors, they also have opportunities to participate in human-machine alignment. Due to AI's autonomy, the objective link between any technician's operational actions and the technology's outcomes is progressively weakened. The prevailing trend is for practitioners to avoid responsibility for the final results of digital technology's intermediary actions. Whether due to subjective or objective weakening of causal responsibility, practitioners should not be overly held accountable for misalignment errors. Users also influence the occurrence and severity of misalignment risks, thereby distributing risk responsibility across more actors. When an actor is both a practitioner and a user, relatively stringent practitioner reviews mitigate their ethical risks in technology use. The primary requirement placed on users is a soft ethical demand, grounded in a sense of community. Therefore, it is necessary to guide users in developing a sense of technological agency, assuming certain moral duties of care and social responsibilities by conveying the relative nature of digital vulnerability and the dynamic characteristics of risk. In fact, it is ordinary citizens with limited understanding of artificial intelligence who are most affected by it, and their voices deserve greater attention.[28]

5.2 Normative Review

The inequality risks emerging in human-machine alignment between individuals essentially represent an expansion of fundamental rights conflicts among private actors, which existing frameworks for rights interpretation and legal application struggle to address. On one hand, if we continue interpreting and reviewing violations of citizens' equality rights solely within the constitutional framework, most infringers who cause substantive harm through digital spaces would evade prosecution. Within the aforementioned tripartite framework, regulating equality rights through a digital lens is an inevitable trend. The determination of equality rights violations will increasingly prioritize outcomes-based assessments of rights infringements over identity-based judgments of actors. Simultaneously, the issue of fundamental rights exerting horizontal effect on private entities becomes more pronounced. Traditional approaches primarily relied on general provisions of private law, which are now inadequate to address the risks outlined earlier. Simultaneously, as conflicts over equality rights among private actors across borders increase, domestic legal frameworks for fundamental rights protection become ineffective. Without private international law or international treaties encompassing

the essence of digital equality rights to provide conflict resolution through applicable law, reinterpretation within existing legal norms becomes necessary—yet such interpretations will grow increasingly strained. German scholars have proposed establishing a digital law with codified characteristics to address the impact of technological development on the logic and framework of fundamental rights.[29] Although the construction of digital law requires rigorous discussion, it offers a normative perspective for protecting equality rights and strengthening anti-discrimination review.

Within this digital law framework, anti-discrimination scrutiny in human-machine alignment requires examining discriminatory discourse or information across all AI interfaces and developing distinct review models. During the initial phase of establishing anti- discrimination mechanisms, proactive review should take precedence. Strengthening proactive review also prevents situations where actual harm has occurred but victims remain unaware of the discrimination and thus fail to assert their rights. Victims' lack of awareness should not serve as a defense against liability for discrimination. Furthermore, judicial review of anti-discrimination requires establishing causation between differential treatment outcomes and protected identity characteristics. AI's intervention in this causal relationship may disrupt traditional reasoning frameworks, necessitating strengthened substantive scrutiny of the link between actions and outcomes. Additionally, anti-discrimination review of decision outcomes requires broader stakeholder participation, creating synergies with subject-based review pathways. Discrimination can be categorized as intent-based or impact-based.[30] Some discriminatory reactions are not intentionally caused by value-bearing entities but arise under AI influence. Therefore, it is necessary to mobilize more entities to participate in anti-discrimination oversight, enabling multiple parties to jointly identify anti-discrimination information and behaviors during backward alignment to calibrate the anti-discrimination value orientation between humans and machines in forward alignment.

5.3 Technical Review

Technology can trigger value conflicts, including those arising from technological practices, conflicts between the intrinsic value of technology and the value of technological harm, and conflicts between the value created by technology and other values. Addressing the legal and social issues involved requires balancing solutions through regulation, beginning with a professional characterization of the technology. [31] Similar to medical technologies subject to ethical review and regulation, artificial intelligence is not value-neutral. It is inherently bound to various value discourses and may destabilize shared human values. Consequently, its safety assessment prioritizes value-based standards. The European Union employs a tiered compliance review and obligation framework based on the risk level AI applications pose to individuals or society as a whole, with most nations adopting similar approaches.

To accommodate technological development, strict scrutiny should not be imposed at the very outset of innovation, and such review is meaningless when technological practices have yet to unfold and value conflicts remain unclear, as it would severely hinder the innovative value of AI technologies. Therefore, a *lenient entry* approach should be adopted for new technological products. The emphasis of review lies on the post-deployment and dissemination phases, where technological value creation that conflicts with equality values should adhere to a *strict exit* review principle. This represents a balancing act between technological development opportunities and uncertain risks and it leads an institutional framework featuring "reserved authorization interfaces.

6.Conclusion

Equality, as a significant achievement and value goal of human civilization, must be a central objective in human-machine alignment. The primary challenge in AI governance lies not in identifying a universal theory of equality, but in establishing design principles to proactively prevent ethical risks. From a normative perspective, whether abstract equality rights or the principle of equal value, only when confronted with specific risks or problems can they be transformed into rights or rules with concrete content.

Historically, human value logic has been subject-centric; however, the integration of AI into moral decision-making demands a paradigm shift. We must move from a purely human communicative model to a subject-interactive paradigm, where value consensus is co-constructed through dynamic interaction between humans and AI. This approach, informed by concepts such as moral reification, integrates non-human entities into the moral community, forming a shared normative order.

Within this new paradigm, the construction of an equality-based value logic must embed technology into the fabric of social development rather than treating it as an external object for static judgment. This requires a predictive process that anticipates the moral impact of technology by incorporating human-AI interaction into our core communicative activities. By doing so, we can challenge the insularity of human-centric ethics, transforming potential ethical confrontation into constructive value alignment. Ultimately, the goal is to establish a dynamic value review mechanism, creating a robust consensus on equality that balances risk prevention with technological innovation. This ensures that humanity guides technology, rather than becoming enslaved by the very tools designed to serve it.

References

[1] Machine Heart Network, "Stuart Russell and Zhu Son gchun's Perspectives on AGI and ChatGPT," February 28, 2023, https://www.jiqizhixin.com/articles/2023-02-28-4?fr

om=synced&keyword=AGI%20and%20ChatGPT E4%B8%8EChatGPT%EF%BC%8CStuart%20Russell%E4%B8%8E%E6%9C%B1%E6%9D%BE%E7%BA%AF%E8%BF%99%E4%B9%88%E7%9C%8B, accessed April 2, 2025.
[2] Toby Ord, *The Precipice: Existential Risk and the Future of Humanity, translated by Wei Silin*, CITIC Press, 2020.
[3] Ji Jiaming, Qiu Tianyi et al., "AI Alignment: A Comprehensive Survey," the AI Safety and Governance Center trans., https://alignmentsurvey.com/uploads/AI-Alignment-A-Comprehensive-Survey-CN.pdf, pp. 1-2, accessed April 5, 2025.
[4] Lu Pingxin, "On Interpretation of the Right to Equality: Structural Reason and Hermeneutic Analysis," *Academic Monthly*, 2022(11).
[5] Hegel, *Encyclopedia of the Philosophical Sciences III: Philosophy of Spirit*, People's Publishing House, 2015.
[6] Eirini Ntoutsi, Pavlos Fafalios et al., "Bias in data-driven artificial intelligence systems—an introductory survey", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2019(3).
[7] Jan Leike, David Krueger et al., "Scalable agent alignment via reward modeling: a research direction", Cornell University, http://arxiv.org/abs/1811.07871.
[8] Yue Yan and Tian Haiping, "Moral Machine and the Moral Prospect of Value Alignment," *Journal of Shenzhen University (Humanities & Social Sciences)*,2024(4).
[9] Tsinghua University Institute for International Governance of Artificial Intelligence website: "Special Forum 4 of the 2022 International Forum on AI Cooperation and Governance—Confronting Gender Discrimination in AI," December 28, 2022, http://aiig.tsinghua.edu.cn/info/1294/1780.htm, accessed March 5, 2025.
[10] International Telecommunication Union: Measuring Digital Development: Facts and Figures 2024, https://www.itu.int/hub/publication/d-ind-ict_mdd-2024-4/, accessed July 30, 2025.
[11] Resolution of the Seventy-Eighth Session of the United Nations General Assembly (A/RES/78/160).
[12] Steven Pinker, *The Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*, New York: Basic Books, 2011.
[13] Ma Changshan, "Identification of Digital Citizens and Protection of Their Rights", *Legal Studies*, 2023(4).
[14] Xu Jing, "On the Connotation, Constitution and Value of the Social Public Power from a Legal Perspective," *China Legal Science,* 2024(1).
[15] Xu Xiaodong and Kuang Yan, "The Generation and Regulation of Algorithmic Power in Governance System," *Journal of Huazhong University of Science and Technology (Social Sciences Edition),* 2024(4).
[16] Cui Jingzi, "Crisis and Response of Equal Rights Protection under the Challenge of Algorithmic Discrimination," *Legal Science*,2019(3).
[17] Amina Adadi and Mohammed Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, 2018(06).
[18] High Level Expert group on Artificial Intelligence in EU, "Ethics guidelines for trustworthy AI," https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
[19] Zhao Yan ed., *Blue Book of Industry and Information Technology: Artificial Intelligence Development Report (2022–2023)*, Social Sciences Academic Press, 2023.
[20] Robert J. Keohane and Joseph Nye, *Power and Interdependence*, Men Honghua trans., Peking University Press, 2012.
[21] Ronald Dworkin, *Taking Rights Seriously*, 1998.
[22] Zhong Haonan, "Digital Power from the Perspective of System Theory, Operational Logic, Alienation Risks and Legal Regulation," *Yunnan Social Sciences*, 2025(1).
[23] Jean-Jacques Rousseau, D*iscourse on the Origin of Inequality*, 1755.
[24] Iason Gabriel, "Artificial intelligence, values, and alignment", *Minds and Machines*, 2020(3).
[25] Yuan Yuqing and Chen Changfeng, "Moralizing Technology: Human-machine Value Alignment and Ethic in Large Language Models", *Nanjing Social Sciences*, 2024(6).
[26] Peter-Paul Verbeek, *Moralizing Technology: Understanding and Designing the Morality of Things,* Chicago: University of Chicago Press, 2011.
[27] Meng Lingyu and Wang Yingchun, "Exploring a New Paradigm for AI Ethical Review", *Science and Society,* 2023(4).
[28] Dai Yibin, "On the Approches to Ethics of Artificial Intelligence", *Social Sciences*,2023(7).
[29] Vaios Karavas and Gunther Teubner, "www.CompanyNameSucks.com：The Horizontal Effect of Fundamental Rights on 'Private Parties' within Autonomous Internet Law"，*German Law Journal*，2003(04)．
[30] Sheila R. Foster, "Causation in antidiscrimination law: Beyond intent versus impact", *Houston Law Review*, 2005(5).
[31] Zheng Yushuang, "Solving the Problem of Technological Neutrality: Jurisprudence Rethinking the Relationship between Law and Technology," *Journal of East China University of Political Science and Law*, 2018(1).