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

Zhang Xinyu
East China University of Political Science and Law

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. Simultaneously, the tension between human-AI interaction brings about new inequality risks under three forms. New equality-value review mechanism aims to improve the safety and trust-worthiness of AI, and additionally grasp opportunities of forging equality-value consensus in risk society.

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.
technological logic may erode and reshape human values and moral sensibilities. Maintaining human-machine value alignment remains imperative. This remains a crucial pathway for mitigating the ethical risks AI development poses to human society.

• Equality is the most vulnerable value. Inequality within human society. The algorithmic black box prevents us from discerning when or in what capacity AI causes inequality, compelling us to guard against value misalignment by prioritizing the greatest predictable risks.

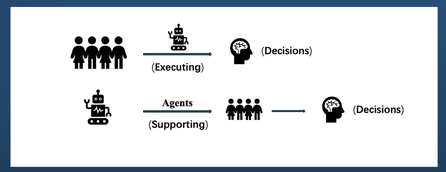
• The meaning of equality, which was originally constantly aligned between people and people is now aligned between humans and artificial intelligence at the same time, and AI currently shows two behavioral roles when participating in ethical decision-making.

• 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.

Purposes

•The 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 techno-logical evolution, and it aims to correspond to broader social changes and to articulate a forward-looking frame-work.

Models



• Figure 1. Two roles of AI in Equality Realization

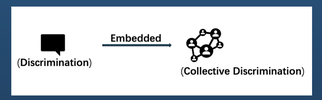
• AI systems centered on large language models primarily learn and mimic human moral directives. Similar to legal representation, large models only operate upon receiving a “mandate” through instructions, strictly executing human morality without exceeding human capacity to generate ethics. To avoid existential risks, AI generates outcomes designed to please.

• AI demonstrates increasingly prominent autonomy in decision-making. 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. 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.

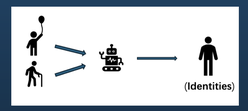
Results

Table 1.
New Characteristics of Traditional Issues

• Through AI’s learning and feedback loops, the speed and outcomes of individual speech can easily outpace the necessary judgment of the speaker. The influence of personal speech in the AI era is highly prone to becoming uncontrollable. 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.

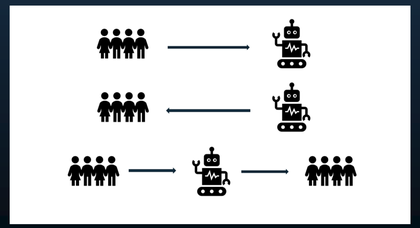


• Traditional discrimination issues not only fail to improve with technological advancement and social progress but become further entrenched。



• Anyone can become the disadvantaged party at any stage of technological development.

Table 2.
New Models of New Inequality Risks



1.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.

2.Technological logic may cause humans to lose certain values and moral senses, though it will also rebuild a form of value consensus. In the interaction of values between humans and ma-chines, artificial intelligence may subvert or replace human rights expression.

3.“tools failing to serve humans”: The cost barrier for accessing AI services has risen, prevent-ng 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.

Discussion

•The primary challenge in AI governance lies not in identifying a universal theory of equality, but in establishing de-sign 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.

• Subject Review: The value actors who first engage with and interact with AI technology are the technical practitioners; Technology users occupy a position between central and peripheral actors, aligning more closely with the former.

•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 le-gal application struggle to address.

•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.

•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.