Can AI Achieve Common Good and Well-being? Implementing the NSTC's R&D Guidelines with a Human-Centered Ethical Approach

Author : Lian Jr-Jiun 連祉鈞

Abstract

This paper delves into the significance and challenges of Artificial Intelligence (AI) ethics and justice in terms of Common Good and Well-being, fairness and non-discrimination, rational public deliberation, and autonomy and control. Initially, the paper establishes the groundwork for subsequent discussions using the Academia Sinica LLM incident and the AI Technology R&D Guidelines of the National Science and Technology Council(NSTC) as a starting point. In terms of justice and ethics in AI, this research investigates whether AI can fulfill human common interests and welfare. Taking AI injustice as an example, I analyze the practical assessment of AI regarding regional, industrial, and social impacts. Further, this paper discusses the challenges of fairness and non-discrimination in AI, specifically addressing the issue of training on biased data, discussing the acquisition of bias by AI and post-processing supervision issues, and emphasizing the importance of rational public deliberation in this process. Then, this research examines the challenges and countermeasures the rational public faces in public deliberation, such as the importance of education in STEM scientific literacy and technological capabilities. Finally, in discussing AI and autonomy, I propose a 'Human-Centered Approach' rather than relying solely on the 'Technological Utility Maximization' brought by AI to achieve substantial AI justice.

Keywords: AI Ethics and Justice, Fairness and Non-Discrimination, Biased Data Training, Public Deliberation, Autonomy, Human-Centered Approach

I. The Academia Sinica LLM Incident and the NSTC's AI R&D Guidelines

Before the release of OpenAI's large language models in October 2022, significant concerns in AI justice and ethics primarily revolved around the application of AI in autonomous vehicles, military drones, and medical diagnostics. However, the launch of OpenAI's ChatGPT large language model(chatbot robot) towards the end of 2022 brought dramatic and impactful changes to academic research in various fields and the general public's perception of AI justice and ethics (Li et al., 2023; Huang et al., 2023; BaHammam et al., 2023; Yi et al., 2023; Wang et al., 2023)..

The societal impact of AI is evident from the recent large language model (LLM) incident at Academia Sinica, which stirred significant social and political concerns about AI in Taiwan. In summary, the incident began when Academia Sinica, on the eve of October 6th (National Day), released an open-source Traditional Chinese large language model developed using Llama 2, CKIP-Llama-2-7b[1-2]. The model's initial presentation as a specialized non-general purpose language model for Ming and Qing dynasty research led to misconceptions of it being a general-purpose Traditional Chinese model, raising greater expectations. Upon widespread use, it was found that the model's responses lacked localization, often using Simplified Chinese terms and politically incorrect responses, sparking widespread debate. Academia Sinica subsequently withdrew the model four days after its release, promising to implement stricter review processes in future research outputs to prevent similar issues. This incident reflects Taiwan's high expectations for localized large language models and highlights the importance of developing such models indigenously.

Researcher Li Yu-Jie(李育杰) from Academia Sinica's Center for IT Innovation noted that owning a large language model is crucial for Taiwan, especially as models provided by OpenAI and Meta exhibit data bias, particularly in collecting Chinese language data. To ensure models are more aligned with local languages and styles, Li emphasized the importance of initial data selection, which must be local and diverse, covering various themes, and should exclude inappropriate terms. Additionally, developing and using LLMs could involve privacy and ethical issues, such as collecting and using personal data. Hence, corresponding policies and laws are needed to regulate and manage the use of such technologies to prevent negative societal impacts[3-8].

Many scholars have already presented research reports on AI's ethical and justice concerns. For instance, Bostrom (2014), in his book " Superintelligence: Paths, dangers, strategies" raises concerns about AI ethics and justice, suggesting that AI surpassing human intelligence could pose threats to humanity. Similarly, Floridi (2014), in "The Ethics of Information," discusses AI ethics and justice, advocating for AI development that aligns with human interests and values.

Taiwan's stance and government measures on AI's political and social issues are as follows: Based on the "Digital Nation and Innovative Economic Development Plan" (2016) announced by NCC and "Taiwan's AI Development Action Plan" (2018) announced by the Executive Yuan, Taiwan has been striving to develop leading AI

infrastructure and a robust ecosystem to create a unique market. Taiwan aims to become a significant global partner in the AI technology and intelligent systems value chain, leveraging hardware and software technology advantages in industries including test fields, regulations, and data-sharing environments. According to the "AI Technology R&D Guidelines" published by the Ministry of Science and Technology (now restructured as the National Science and Technology Council) in September 2019, the government considers AI technology development a critical direction and a cross-ministerial issue requiring long-term investment and international alignment[9].

II、Can AI Achieve Common Good and Well-being? On AI Epistemic Injustice

Some scholars and the public may be naively optimistic about the development of AI technology, believing that the era of AI will shape a world where "knowledge is no longer the patent of specific groups, knowledge knows no borders, and sharing is boundless" thereby promoting the Common Good and Well-being of humanity as a whole (see Su, 2023). However, this optimistic attitude may need to be revised.

Taking the incident of Academia Sinica's large language model(LLM) in Taiwan in 2023 as an example, constrained by a budget of only three hundred thousand NTD[2], the trained Traditional Chinese large language model CKIP-Llama-2-7b suffered from an inadequate corpus, ultimately leading to its controversial withdrawal from use; in contrast to ChatGPT's large language model(LLM), which was supported by a training, server, and maintenance budget exceeding one hundred billion U.S. dollars[10]. There is a visible 'benefits and welfare' gap between CKIP-Llama-2-7b and ChatGPT; one of the main causes is the 'AI Injustice' arising from asymmetries in information, corpus databases, funds, and technology. The basic requirements for large language models and deep machine learning include powerful GPUs for computation, high-performance HGX platform servers, and large corpora for training, necessitating a substantial engineering workforce for training and adjustments (Dettmers et al., 2022; Gururangan et al., 2023; Narayanan et al., 2021). Thus, the original expectation of some scholars that AI technology could democratize knowledge may significantly differ from reality: "Knowledge and technology will serve those with more resources." This outcome evidently fails to achieve the vision of 'knowledge without borders, shared without limitations.' Following this logic, the same 'AI Injustice' will inevitably occur among entities with asymmetrical resources (especially in technology and funding): between nations, governments and corporations, and corporations and individuals. For instance, some countries or organizations may have access to large datasets and advanced technology, while others may lack these resources.

The rapid development of AI technology brings opportunities but also a host of ethical and justice issues. When discussing the shaping of an era where "knowledge is no longer the patent of specific groups, knowledge knows no borders, and sharing is boundless" through AI, we must recognize that this optimistic expectation may be challenged by asymmetries in resources stemming from social and structural power

imbalances, which have been widely discussed in AI ethics literature (Birhane et al., 2022). The Academia Sinica CKIP-Llama-2-7b incident is a clear example of resource asymmetry. When a country or organization's funding is limited, the development and application of its AI technology are likely to be restricted. This asymmetry in resources reflects not only in financial capital but also in technology, data, and talent. When a country, organization, or research unit has limited funding, the development and application of its AI technology are likely to be constrained. This asymmetry in resources is not only about pure capital investment but also spans across interdisciplinary technology, data corpora, and diverse talent pools. Some advanced countries or corporate organizations may possess larger datasets and more advanced technology. In contrast, others may lack these resources (Leavy & O'Sullivan, 2020), leading to an increasing gap in resource asymmetry.

Furthermore, when we expect AI technology to promote 'Common Good and Well-being,' we must also recognize that certain groups or individuals might monopolize this technology. This means that while AI technology has the potential to facilitate global knowledge sharing, in practice, it may be controlled and utilized by resource-rich groups. Thus, when discussing ethical and justice issues of AI, we must consider these potential risks and challenges. Only through equitable and transparent means can we ensure that AI technology genuinely serves humanity's Common Good and Well-being (Taiwo et al., 2023).

The rapid advancement of AI technology has not only led to breakthroughs on a technological front but has also given rise to a series of ethical and justice-related issues. These issues are often associated with resource asymmetry, technology fairness, and how specific groups control and exploit technology. When we hope that AI technology can foster an era where "knowledge is no longer proprietary to certain groups, knowledge knows no borders, and sharing is limitless," we must delve deeply into these ethical and justice issues. Firstly, the development and application of AI technology are subject to resource constraints, encompassing not just financial aspects but also technology, data, and expertise. Such resource asymmetry could restrict technological development and application, thereby impacting the fairness and justice of technology (Leavy & O'Sullivan, 2020). Moreover, resource asymmetry may lead to the control and exploitation of technology by specific groups or individuals, posing a severe challenge to the fairness and justice of technology (Birhane et al., 2022). Secondly, the development and application of AI technology are also influenced by ethical considerations. In recent years, scholars and experts have explored the ethical issues surrounding AI technology and have proposed a range of guiding principles and recommendations (Mittelstadt, 2019). However, these guidelines and recommendations often lack practical operability and may be limited by resources and controlled by specific groups in their actual application (Pant et al., 2022).

III. Can AI Achieve Common Good and Well-being? Social Impact Assessment

Overall, to ensure that AI technology truly serves the Common Good and Well-being of all humanity, from the perspective of AI justice, we can approach this from a more practical standpoint:

1.Regional Variations in AI Technology:

The development of AI technology demonstrates significant regional differences. For example, the United States and China have taken leading positions in the research and application of AI technology, thanks to their robust economic foundations, wealth of technical talent, and substantial government support (Chan et al., 2021). In contrast, Africa and some Southeast Asian countries are lagging in AI development, primarily due to insufficient funding, technical talent, and infrastructure (Blasi et al., 2021). Such regional technological disparities not only affect the global competitive standing of nations but may also exacerbate international technological inequalities (Hagerty & Rubinov, 2019). With globalization, the differences in AI development and application between regions are increasingly attracting international attention.

These regional disparities reflect the competitive capacity of nations in technological development and reveal the reality of global technological inequality. To narrow this gap, many countries and international organizations are seeking opportunities for technology transfer and collaboration. For instance, the United Nations Development Programme (UNDP) has initiated several AI technology cooperation projects aimed at assisting developing countries in improving their AI research and application capabilities. However, these technology transfers and collaborations also face many challenges, such as intellectual property protection, technology standardization, and talent training (Park, 2023). The regional disparities in AI technology are also closely related to the economic development levels of countries. On one hand, technological development can promote economic growth and innovation, enhancing national competitiveness; on the other hand, technological disparities may exacerbate economic inequalities, placing greater pressure on disadvantaged nations and regions (Li et al., 2023).

To reduce regional disparities in AI technology, education and training are key factors. Many countries have recognized this and are investing in AI education and training programs to cultivate more technical talent. Closing the regional gap in AI technology requires the collective effort of the international community. Nations should strengthen cooperation, share technology and experience, and provide technical assistance and financial support to developing countries. Additionally, international organizations such as the United Nations, the World Bank, and the International Monetary Fund should also play their roles in promoting the international equity, justice, and sustainability of AI technology. (Kuhlman et al., 2020; Madaio et al., 2021; Ho et al., 2023; Viberg et al., 2023; Kacperski et al., 2023).

2. Industrial Applications of AI Technology:

At the industrial level, the application of AI technology also presents different competitive landscapes. For example, the finance and healthcare industries are the most widespread domains of AI applications, with leading companies like JPMorgan Chase and Siemens Healthineers optimizing business processes, enhancing service quality, and innovating products through AI (Andreu-Perez et al., 2018). However, the agricultural and manufacturing sectors are relatively behind in AI application, mainly due to a mismatch between the characteristics and demands of these industries and AI application scenarios (Lu et al., 2023; Nelson et al., 2023). The application of AI technology in various industries depends not only on the maturity of the technology but also on industry characteristics, market demand, and the policy environment.

There are further explore the application and challenges of AI technology in different industries. For instance,

(i) Finance: the finance sector is an early adopter of AI, particularly in risk assessment, investment strategy, and customer service. Many banks now use AI for credit scoring to assess customer credit risk more accurately (Adekunle et al., 2023). Moreover, AI is employed in high-frequency trading and asset management, aiding investors in making better decisions (Maple et al., 2023; Rasouli et al., 2023).

(ii) Healthcare: AI's application in healthcare is widespread, from disease diagnosis to drug development(Ting et al., 2018). Google's DeepMind, for example, has developed an AI algorithm capable of quickly and accurately diagnosing eye diseases, significantly improving diagnostic accuracy and efficiency(Cheung et al., 2019). AI is also utilized in drug development, predicting new drugs' effects and side effects by analyzing vast amounts of biological data(Mak&Pichika, 2019; Paul et al., 2021).

(iii) Agriculture: Although lagging in AI application, agriculture has seen interesting use cases emerge recently. Agricultural robots, for instance, can use AI for automated crop picking and sorting, greatly improving production efficiency (Martini et al., 2022). AI can also monitor farmland by analyzing satellite imagery to predict crop growth and pest outbreaks. (Victor et al., 2022; Sykas et al., 2022; Bassine et al., 2023).

(iv) Manufacturing: the manufacturing industry also faces challenges in AI application but has seen success stories. Manufacturers can achieve intelligent production by using AI to automatically adjust production line operations to meet market demand (Ong et al., 2020).

AI can also predict equipment failures, allowing for preemptive maintenance and reducing the risk of production downtime (Ge et al., 2022;

Dosluoglu & MacDonald, 2022; Mangal & Kumar, 2016; Nelson et al., 2023).

In summary, AI technology has progressed in various industries but also faces challenges. These challenges relate to technological maturity, industry characteristics, market demand, and the policy environment. From an industrial perspective, to fully unleash the potential of AI, each sector needs to collaborate deeply with technology providers, research institutions, government departments, and relevant think tanks to promote the development and application of AI collectively.

3. Socio-Cultural Impact of AI Technology:

The widespread adoption and application of AI technology have altered the competitive dynamics in the economy and industry and profoundly impacted society and culture. For example, applying AI puts many traditional job roles at risk of displacement while simultaneously creating many new employment opportunities. AI has also revolutionized sectors like education, healthcare, and entertainment, making services and products in these areas more intelligently adaptable and personalized. The societal impact of AI is multifaceted and continues to grow in depth and scope (Venkatasubramanian et al., 2020; Nelson et al., 2023; Tucker et al., 2020; Dignum, 2022; Hagerty & Rubinov, 2019). Here, we will briefly explore the impact of AI in different social domains and their associated challenges and opportunities. Firstly, (i) labor market transformation is one of AI's most direct and noticeable societal impacts. Traditional, repetitive, and low-skill jobs, such as customer service, data entry, and basic manufacturing tasks, face the risk of automation (Chui et al., 2016). However, this also means that more opportunities are emerging in high-skill and creative areas, such as AI algorithm design, machine learning research, and strategic planning for artificial intelligence (Brynjolfsson & McAfee, 2017). This labor market transformation requires workers to continually learn and adapt to meet the challenges posed by technology. Secondly, (ii) the field of education is also deeply affected by AI. Traditional teaching models are being replaced by intelligent tutoring systems and personalized learning platforms that can adjust based on a student's progress and interests, offering a more personalized learning experience (Woolf, 2010). Moreover, AI can assist educators in analyzing students' learning data, predicting learning difficulties, and providing timely assistance and support (Baker & Siemens, 2014). In the (iii) healthcare sector, AI is changing how diseases are diagnosed and treated. For instance, deep learning techniques can help doctors diagnose diseases like cancer, diabetes, and heart disease more quickly and accurately; AI is also used in drug development, predicting the effects and side effects of new medications by analyzing large sets of biological data (Esteva et al., 2017). The (iv) entertainment industry is also impacted by AI. For example, AI can be used to create music, films, and games, providing

richer and more varied entertainment experiences. It can also enhance user experience by recommending content tailored to individual preferences through algorithmic curation (Goodfellow et al., 2016). However, the proliferation and application of AI also raise a series of social and ethical issues, such as data privacy and security, technological unemployment, and biases and unfairness in AI decision-making (Crawford & Calo, 2016). These issues require the collective effort of governments, businesses, and societies to develop appropriate policies and laws to ensure the healthy and sustainable development of AI technology. The spread and application of AI are profoundly changing our society and culture, bringing many opportunities and challenges.

III. Fairness and Non-Discrimination: Examining the Importance of Public Deliberation in Training with Biased Data"

According to the National Science and Technology Council's "Guidelines for Research and Development(R&D) in Artificial Intelligence" [9], the principle of "Fairness and Non-discrimination" states that "researchers should strive to ensure that AI systems, software, algorithms, and other technologies, as well as their decision-making processes, are Human-Centered, equally respecting the fundamental human rights and dignity of all individuals, avoiding risks of bias and discrimination, and establishing external feedback mechanisms." Under this guideline, it is crucial to consider whether AI systems may generate biases and discrimination that violate the spirit of Human-Centered values.

As previously mentioned, the rapid development and application of AI technology, playing a pivotal role in various fields from financial forecasting and medical diagnosis to social media recommendations, has raised increasing concerns about potential biases and discrimination in decision-making processes (O'Neil, 2016). Firstly, AI systems' decision-making often relies on extensive data, which may originate from historical records or user behavior. However, if these data inherently contain biases, AI systems might 'learn' these biases during training, subsequently reproducing them in future decisions (Barocas & Selbst, 2016). For example, if past recruitment data shows a lower hiring rate for certain genders or ethnic groups, an AI recruitment system trained on this data might be biased against recommending candidates from these groups. Secondly, even if the data itself is unbiased, biases can still arise in AI systems during data processing. This is because AI models might overfit certain data features during training, thereby overlooking other crucial information (Zhang et al., 2021). Such overfitting can lead to unfair decision-making in practical applications. To address these issues, researchers and engineers have developed various methods to enhance the fairness of AI systems and reduce discrimination. One approach is the use of 'fairness constraints,' which involves incorporating specific justice and ethical fairness constraints during the model training process to ensure that the model's decisions do not adversely affect certain groups (Zafar et al., 2017). Another method is 'post-processing,' which adjusts the decision outcomes of models after training to ensure fairness (Hardt et al., 2016). Besides

technical methods, increasing the transparency and explainability of AI systems is also considered key to addressing bias and discrimination (Doshi-Velez & Kim, 2017). If users and decision-makers can clearly understand the decision-making processes and bases of AI systems, they are better equipped to identify and correct potential biases.

The decision-making process of AI can be influenced by its training data, leading to biases and discrimination. This not only potentially results in unfair treatment of certain groups but also poses threats to the harmony and stability of society (Buolamwini & Gebru, 2018). For instance, when AI technology is applied in medical diagnosis, if its training data primarily consists of medical records from a specific group, the AI system might produce inaccurate or even erroneous diagnoses for another specific group (Obermeyer et al., 2019). This could lead to not only a waste of overall governmental medical resources but also severe threats to the health and safety of many patients. Additionally, the application of AI technology in the financial sector, such as credit scoring and loan approval, might also result in unfair treatment of certain groups due to biases in its inherent algorithms or training samples. For example, if an AI credit scoring system's training data mainly consists of credit records from high-income groups, the system might be overly conservative in assessing the credit risk of low-income groups, thereby rejecting their loan applications. These issues are related not only to the limitations of AI technology itself and costs (resources and funds) but also to the quality and diversity of its training data(Chouldechova & Roth, 2018). Therefore, to ensure the fairness and non-discrimination of AI technology, efforts need to be made in multiple aspects. Firstly, more diverse and comprehensive training data must be collected to ensure that AI systems can better understand, represent, and reflect the diversity of the real world (Gebru et al., 2018). Secondly, more advanced AI models and algorithms need to be developed to reduce inherent algorithmic biases and discrimination (Kearns & Roth, 2019). Lastly, human oversight and review of AI systems must be strengthened to ensure their fairness and non-discrimination in practical applications (Raji et al., 2020). These points indicate that the challenges are not only related to the technology itself but also closely linked to how we should consider the ethical and justice implications of AI applications.

When discussing the ethics and justice of AI, it is imperative to consider how technology affects people's lives and rights. For example, AI technology might exacerbate social inequalities, benefiting some groups while harming others, potentially leading to societal division and conflict, and posing threats to social harmony and stability (Jobin, Ienca, & Vayena, 2019). Therefore, it is necessary to ensure that the development and application of AI technology are just and ethical, respecting the rights and dignity of every individual. To achieve this goal, efforts must be made in multiple areas. Firstly, ethical and justice education regarding AI technology must be strengthened, enabling researchers and engineers to understand and recognize the potential risks and challenges of AI technology. Secondly, a set of just and transparent guiding principles and standards must be established to guide the development and application of AI technology (Taiwo et al.,

2023). These principles and standards should be based on human rights and dignity, ensuring that the development and application of AI technology do not infringe upon these rights. Thirdly, the regulation and review of AI technology must be strengthened to ensure its fairness and non-discrimination in practical applications. This includes establishing a just and transparent evaluation and review mechanism to assess the potential risks and challenges of AI technology and ensuring that it does not infringe upon human rights and dignity. Fourthly, rational public participation and social supervision must be strengthened, enabling the public to participate in the decision-making process of the development and application of AI technology, ensuring that it aligns with societal values and expectations.

Rational public participation plays a crucial role in the development and application of AI technology. Not only can it ensure that the development and application of AI technology are more just and ethical, but it can also enhance public trust and acceptance of AI technology. When the rational public participates in the decision-making process of AI technology, they can provide valuable opinions and suggestions, helping researchers and engineers better understand and address the potential risks and challenges of AI technology. Additionally, rational public participation can provide a platform for different stakeholders to discuss and negotiate. This underscores the importance of public participation and deliberation in discussions about the ethics and justice of AI. Rational public involvement ensures that the development and application of AI technology are more transparent and align more closely with societal values and expectations (Eubanks, 2018). Additionally, it provides a platform for various stakeholders to collectively discuss and decide on the direction and strategies for AI technology development and application (Hagendorff, 2020). This ensures that the development and application of AI technology are more just and ethical and strengthens public trust and acceptance of AI technology. Rational public participation plays a key role in developing and applying AI technology. Firstly, it offers a deliberative platform for the public to understand AI technology's potential risks and challenges and provide their opinions and suggestions. This ensures that the development and application of AI technology align with public needs and expectations, and mitigates potential risks and challenges. Secondly, rational public participation also enhances public trust and acceptance of AI technology. When the public is involved in the decision-making process of AI technology development and application, they are more likely to trust and accept AI technology and are more willing to use and support it (Dignum, 2019).

IV. AI Justice and Rational Public Deliberative Participation

The development and application of AI technology involve not only technical issues but also social, cultural, and political aspects. In this context, public deliberative participation becomes a key factor in ensuring the fairness and non-discrimination of AI technology (Fraser, 1990). Public deliberation provides diverse perspectives and opinions and enhances the social legitimacy and acceptance of AI technology (Gutmann & Thompson,

2004). In today's technological era, AI has permeated various aspects of our lives, including healthcare, transportation, and finance. However, as AI becomes more widespread, the ethical and justice issues it raises are increasingly attracting societal attention (Bryson, 2018). Against this backdrop, public deliberation is crucial in guiding the direction of AI technology, ensuring that its development is a technological advancement and a social and moral progression (Latour, 2004).

The core concept of public deliberation is the 'rational public,' which involves open and rational discussions under conditions of equality and justice to reach collective decisions (Habermas, 1984). In the development and application of AI technology, public deliberation ensures that the direction and strategies of technology align with societal values and expectations, avoiding potential biases and discrimination (Young, 2000). For example, when AI is applied in urban planning and management, public participation in decision-making can ensure that AI applications are more just and reasonable, reflecting the public's expectations and needs for urban development (Fung, 2006). However, rational public deliberation in AI development and application faces many challenges. Firstly, the public may lack knowledge and understanding of AI technology, making it difficult for them to participate in and make decisions about its development and application (Jasanoff, 2003; Mittelstadt et al., 2016). Moreover, the process of public deliberation may be influenced and manipulated by certain stakeholders or vested interests, potentially leading to the neglect or distortion of public opinions and suggestions (Cath et al., 2018; Mansbridge et al., 2012).

Therefore, it is necessary to ensure that the process of rational public deliberation is just and transparent, providing sufficient resources and support for the public to effectively participate in and make decisions about the development and application of AI technology (Fishkin, 2009). To ensure the effectiveness of public deliberation, governments need to make efforts in several areas. Firstly, governments and educational institutions should strengthen public education and training in AI technology, enabling the public to understand its basic knowledge and principles and critically think about and assess its potential risks and challenges (Dahlberg, 2001). This can help the public participate more effectively in decision-making about AI development and application. Secondly, governments need to establish just and transparent mechanisms for public deliberation, ensuring that public opinions and suggestions are fully considered and reflected in the development and application of AI technology (Benhabib, 1996; Bietti, 2020). Thirdly, governments and policy implementers need to enhance communication and interaction with the public, providing objective indicators and suggestions for reference (Crawford & Calo, 2016), ensuring that the public fully understands the relevance and importance of AI technology development and application to their interests and national development (Dryzek, 2000).

In the academic realm, we closely monitor the rapid development of AI technology, highlighting the growing importance of rational public deliberative

participation. However, as AI is applied in various fields such as medical diagnosis, financial transactions, and traffic management, its decision-making processes and outcomes significantly impact the public's interests and rights (Bostrom, 2014). Yet, this impact may not be immediately apparent to the general public. Therefore, it is essential to ensure that the public has the right to participate in these decision-making processes, guaranteeing that the application of AI technology is fair and just (Vallor, 2016).To achieve this goal, it is crucial first to ensure that the public possesses sufficient knowledge and capability to engage in these discussions. As highlighted in this paper, a major challenge in public deliberative participation is ensuring that the participating public has enough knowledge and capability to engage in the discussion. The public's knowledge and capabilities refer not only to a basic understanding of AI technology but more importantly, to their ability to think critically, assess the potential risks and challenges of AI technology, and propose concrete suggestions and solutions (Turkle, 2015). For this purpose, governments need to strengthen public education, especially in STEM (Science, Technology, Engineering, and Mathematics) fields, to cultivate the public's scientific literacy and technical abilities (Resnick & Wilensky, 1998). Additionally, governments need to provide more resources and platforms for the public to participate in the research and development of AI technology, thereby ensuring their right to speak during the deliberation process (O'Neil, 2016).

Besides enhancing the public's knowledge and capabilities, it is also necessary for governments to ensure the fairness and transparency of the deliberation process. This means establishing a just and transparent deliberative mechanism, ensuring that all rational participants can engage in discussions on equal terms and that their opinions and suggestions are fully considered (Rawls, 1971). Moreover, it is crucial to ensure that the results of public deliberation are effectively implemented. This requires establishing an effective implementation mechanism to ensure that the results of deliberation, as a form of "communicative action," are translated into practical policies and measures (Habermas, 1996). While many citizens in the academic community are well-informed and rational, not all well-educated and rational citizens can fully understand and evaluate the potential risks and challenges of AI technology (Anderson et al., 2021). Therefore, a system is needed to improve the AI literacy of the rational public through education and training, enabling them to participate more effectively in the deliberative process (Brey, 2012). Furthermore, public deliberation faces the challenge of ensuring the fairness and transparency of the participation process. In today's society, certain stakeholders may attempt to manipulate the deliberative process to maximize their interests (Fricker, 2021). To prevent this, a just and transparent deliberative mechanism must be established, ensuring that all participants can engage in discussions on equal terms and that their opinions and suggestions are fully considered (Gastil & Levine, 2005).

Further, we recognize the challenge in ensuring the effective implementation of the outcomes of public deliberation. Public deliberation results often involve multiple interests, and conflicts between these interests are common, such as in urban

redevelopment, the placement of mass transit systems, and the externalization of public construction costs (NIMBY effects). It's not always possible for decisions made through public deliberation to gain universal support and agreement (Mouffe, 2013). Therefore, governments need to establish more effective implementation mechanisms to ensure that the results of deliberation are translated into practical policies and measures, especially benefiting the most vulnerable populations (Niemeyer, 2014; Rawls, 1971).In light of this, it's evident that ensuring the effectiveness of public deliberative participation involves overcoming many challenges. These range from improving public technical literacy to establishing just and transparent deliberative mechanisms. Efforts are required from government entities, civil organizations, and rational citizens themselves to make systemic changes (Sunstein, 2017). Only through these efforts can we ensure that the development and application of AI technology are fair and just, gaining the support and recognition of the rational public (Floridi & Taddeo, 2016). This discussion underscores the crucial role of rational public participation in the development and application of AI technology. Governments need to ensure that the process of rational public participation is just and transparent, providing sufficient resources and support for the public to effectively participate in and make decisions about AI technology's development and application. Ultimately, this will ensure that the development and application of AI technology are just, ethically fair, and institutionally respectful of everyone's equal rights and autonomous dignity as citizens.

V. AI and Autonomy: Human-Centered Ethical Approach

The National Science and Technology Council's "Guidelines for Research and Development in Artificial Intelligence"[9] emphasize the principle of "Fairness and Non-discrimination," stating that "AI applications should assist human decision-making, and AI researchers should focus on enabling humans to retain complete and effective autonomy and control over these systems." However, the extent of human decision-making power in the recent development of AI technology faces significant challenges and impacts. For instance, human control over braking and lane changing in fully autonomous vehicles seems nearly nullified. Similarly, Taiwan's recent initiative to assist judges in drafting court decisions using AI raises questions about human autonomy and control in such scenarios. Understanding 'autonomy and control' in operational terms is crucial. Autonomy is typically understood as the ability of individuals to make choices freely based on their will and beliefs. At the same time, control refers to their ability to manage these choices and actions (Dworkin, 1988). In the context of AI, this implies that humans should be able to decide when and how to use AI technology and control its outputs and outcomes (Turkle, 2011).

In this digital and automated age, the pace of AI technology development far exceeds our expectations, posing unprecedented challenges to human autonomy and control in various domains (Bryson, 2016). From autonomous vehicles to medical diagnostics and financial transactions, AI technology gradually replaces human decision-

making roles, threatening human autonomy and control (Scherer, 2016). We must recognize that the development of AI technology is not just a technological advancement but also a social and ethical challenge. The application of AI in various fields brings efficiency improvements and significant impacts on human autonomy and control (Vallor, 2016). For example, AI in medical diagnostics could challenge doctors' professional judgments, shifting treatment decisions from doctors to AI systems (Char, 2018). Moreover, the development of AI raises numerous ethical and moral issues, such as its application in the judicial system potentially affecting the fairness and justice of legal decisions (Dressel & Farid, 2018). When AI technology is applied in financial trading, it could potentially shift decision-making from humans to AI systems, posing a serious threat to the fairness and transparency of financial markets (Chui & Zeng, 2016). Thus, we must recognize that the development of AI technology is not just a technological advancement but also a social and ethical challenge. It is imperative to ensure that AI development occurs within a moral and ethical framework, safeguarding human autonomy and control (Bostrom, 2014). This necessitates establishing a comprehensive legal and policy framework to ensure that AI development adheres to principles of fairness and justice (Cath et al., 2018), protecting our inherent and cherished human autonomy and control.

However, as AI technology advances, this autonomy and control seem to be gradually eroding. For instance, in medical diagnostics, patients might rely entirely on AI recommendations, overlooking their bodily sensations and intuition (Topol, 2019). Similarly, investors might depend solely on AI advice in financial trading, ignoring market dynamics and risks (Pasquale, 2015). This over-reliance on AI technology could lead to a loss of human autonomy and control in many critical decisions, posing not just a technological issue but also a moral and ethical one. If humans lose autonomy and control, AI technology could entirely dominate our lives and future, threatening our freedom and dignity (Harari, 2018). To address this, we need to rethink the direction and goals of AI development. In this digital age, expectations of technology are increasing, while expectations of each other seem to be decreasing. What lies behind this phenomenon? Turkle (2011) points out that with technological advancement, people increasingly rely on machines for various tasks, reducing human-to-human interaction. This affects social skills and poses serious challenges to human autonomy and control. Over-reliance on technology can lead to losing control over our lives, such as excessive dependence on smartphones for information and communication, potentially leading to social alienation (Przybylski & Weinstein, 2013). Moreover, over-reliance on AI for decision-making could threaten human autonomy and control (O'Neil, 2016). This phenomenon also reflects a significant trend in contemporary society: the pace of technological development far exceeds human adaptability, often leaving people reliant on technology for various tasks (Carr, 2010).

The issue at hand extends beyond mere technological concerns; it fundamentally pertains to moral and ethical dimensions. The potential loss of human autonomy and

control to technology poses a profound threat to our freedom and dignity, as highlighted by Zuboff (2019). This situation underscores the necessity of adhering to a Human-Centered approach in technological development, emphasizing that "technology originates from human nature." Such an approach is crucial to prevent the technological erosion of our intrinsic freedoms and autonomous dignity.In the rapid development of AI technology, we need to ensure that its development is "Human-Centered," not "technology-utility-centric." This means enhancing human welfare through AI, not merely pursuing technological utility maximization(see Floridi & Cowls, 2019; Morozov, 2013). Additionally, we must ensure that AI development occurs under principles of fairness and justice. This involves ensuring that AI applications do not lead to social inequality and discrimination and that everyone can benefit equitably from AI advancements (Eubanks, 2018; Mittelstadt & Floridi, 2016). Therefore, the questions and challenges surrounding "AI ethics and justice" and "autonomy and control" are complex and urgent. We need in-depth research and discussion on this issue and effective measures to ensure that AI development occurs under principles of fairness and justice, safeguarding human autonomy and control against the threats posed by rapid technological advancement (see Vinuesa et al., 2020; Winner, 2010).

IV. Conclusion

In conclusion, this paper has delved into the justice and democratic aspects of the National Science and Technology Council's AI Research and Development Guidelines. It began by examining the issues of epistemic injustice and social injustice in AI, exploring the question of whether AI can achieve the common good and well-being. Subsequently, through a discussion on 'Fairness and Non-Discrimination,' the paper highlighted the significance of Rational Public Deliberative Participation in correcting biased data. Finally, in addressing the issue of autonomy, a call was made for a 'Human-Centered Approach' as a means to achieve substantial AI justice. This comprehensive analysis underscores the multifaceted challenges posed by AI technology, emphasizing the need for a balanced approach that considers ethical, social, and democratic dimensions. By advocating for a human-centered approach, the paper stresses the importance of maintaining human autonomy and control in the face of rapidly advancing AI technologies. It also points out the critical role of public participation and deliberation in shaping AI development, ensuring that it aligns with societal values and addresses issues of fairness and bias. The paper's exploration of these themes contributes to the ongoing discourse on AI ethics and governance, offering insights into how we can steer AI development towards benefiting society as a whole. It calls for collaborative efforts among policymakers, technologists, ethicists, and the public to create AI systems that are not only technologically advanced but also socially responsible and ethically sound.

Reference

Adekunle, O., Riedl, A., & Dumontier, M. (2023). Models towards Risk Behavior Prediction and Analysis: A Netherlands Case study. *arXiv preprint arXiv:2311.04164*.

Anderson, C., Brabham, D., & Feller, A. (2021). Crowdsourcing the Public Participation Process for Planning Projects. *Planning Theory*, *8*(3), 242-262.

Andreu-Perez, J., Deligianni, F., Ravi, D., & Yang, G. Z. (2018). Artificial intelligence and robotics. *arXiv preprint arXiv:1803.10813*.

BaHammam, A. S., Trabelsi, K., Pandi-Perumal, S. R., & Jahrami, H. (2023). Adapting to the Impact of AI in Scientific Writing: Balancing Benefits and Drawbacks while Developing Policies and Regulations. *arXiv preprint arXiv:2306.06699*.

Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 379-396.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California law review*, 671-732.

Bassine, F. Z., Epule, T. E., Kechchour, A., & Chehbouni, A. (2023). Recent applications of machine learning, remote sensing, and iot approaches in yield prediction: a critical review. *arXiv preprint arXiv:2306.04566*.

Benhabib, S. (1996). *Democracy and difference: Contesting the boundaries of the political*. Princeton University Press.

Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. *Philosophy & Technology*, 1-20.

Birhane, A., Ruane, E., Laurent, T., Brown, M. S., Flowers, J., Ventresque, A., & Dancy, C. L. (2022). The Forgotten Margins of AI Ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 948-958).

Blasi, D., Anastasopoulos, A., & Neubig, G. (2021). Systematic Inequalities in Language Technology Performance across the World's Languages. *arXiv preprint arXiv:2110.06733*.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. *Oxford University Press*.

Brey, P. (2012). Anticipating ethical issues in emerging IT. Ethics and Information Technology, 14(4), 305-317.

Brynjolfsson, E., & McAfee, A. (2017). The business of artificial intelligence. Harvard Business Review.

Bryson, J. (2016). AI and society: understanding the human use of AI. *Neural Networks*, 100, 59-64.

Bryson, J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. Ethics and Information Technology, 20(1), 15-26.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1-15.

Carr, N. (2010). *The shallows: How the internet is changing the way we think, read and remember*. Atlantic Books Ltd.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science and engineering ethics*, 24(2), 505-528.

Chan, A., Okolo, C. T., Terner, Z., & Wang, A. (2021). The limits of global inclusion in AI development. *arXiv preprint arXiv:2102.01265*.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine*, 378(11), 981-983.

Cheung, C. Y., Tang, F., Ting, D. S. W., Tan, G. S. W., & Wong, T. Y. (2019). Artificial intelligence in diabetic eye disease screening. *The Asia-Pacific Journal of Ophthalmology*, *8*(2), 158-164.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810.

Chui, M., & Zeng, M. (2016). The rise of the machines. *Foreign Affairs*, 95, 16.

Chui, M., Manyika, J., & Miremadi, M. (2016). Where machines could replace humans—and where they can't (yet). *McKinsey Quarterly*.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature News*, 538(7625), 311.

Dahlberg, L. (2001). The internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. Information, Communication & Society, 4(4), 615-633.

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer Nature.

Dignum, V. (2022). Relational artificial intelligence. *arXiv preprint arXiv:2202.07446*.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dosluoglu, T., & MacDonald, M. (2022). Circuit Design for Predictive Maintenance. *arXiv preprint arXiv:2211.10248*.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1), eaao5580.

Dryzek, J. S. (2000). Deliberative democracy and beyond: Liberals, critics, contestations. *Oxford University Press*.

Dworkin, G. (1988). *The theory and practice of autonomy*. Cambridge University Press.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. *St. Martin's Press*.

Fishkin, J. S. (2009). *When the people speak: Deliberative democracy and public consultation*. Oxford University Press.

Floridi, L. (2014). *The ethics of information*. Oxford University Press.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).

Floridi, L., & Taddeo, M. (2016). What is data ethics?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160360.

Fraser, N. (1990). Rethinking the public sphere: A contribution to the critique of actually existing democracy. Social text, (25/26), 56-80.

Fricker, M. (2021). Evolving Concepts of Epistemic Injustice. *International Journal of Political Philosophy/Las Torres de Lucca*, *10*(19).

Fung, A. (2006). Varieties of participation in complex governance. Public administration review, 66(s1), 66-75.

Gastil, J., & Levine, P. (2005). The deliberative democracy handbook: Strategies for effective civic engagement in the twenty-first century. Jossey-Bass.

Ge, N., Li, G., Zhang, L., & Liu, Y. (2022). Failure prediction in production line based on federated learning: an empirical study. *Journal of Intelligent Manufacturing*, *33*(8), 2277-2294.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). *Cambridge: MIT press*.

Gururangan, S., Li, M., Lewis, M., Shi, W., Althoff, T., Smith, N. A., & Zettlemoyer, L. (2023). Scaling Expert Language Models with Unsupervised Domain Discovery. *arXiv preprint arXiv:2303.14177*.

Gutmann, A., & Thompson, D. (2004). Why deliberative democracy? *Princeton University Press*.

Habermas, J. (1984). The theory of communicative action: Reason and the rationalization of society (Vol. 1). *Beacon press*.

Habermas, J. (1996). The European nation state. Its achievements and its limitations. On the past and future of sovereignty and citizenship. *Ratio juris*, *9*(2), 125-137.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99-120.

Hagerty, A., & Rubinov, I. (2019). Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv preprint arXiv:1907.07892.

Hardt, M., Megiddo, N., Papadimitriou, C., & Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*(pp. 111-122).

Harari, Y. N. (2018). 21 lessons for the 21st century. *Spiegel & Grau*.

Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A., ... & Snidal, D. (2023). International institutions for advanced AI. arXiv preprint arXiv:2307.04699.

Huang, Y., & Xiong, D. (2023). Cbbq: A chinese bias benchmark dataset curated with human-ai collaboration for large language models. *arXiv preprint arXiv:2306.16244*.

Jasanoff, S. (2003). Technologies of humility: Citizen participation in governing science. *Minerva*, 41(3), 223-244.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, *1*(9), 389-399.

Kacperski, C., Ulloa, R., Bonnay, D., Kulshrestha, J., Selb, P., & Spitz, A. (2023). Who are the users of ChatGPT? Implications for the digital divide from web tracking data. *arXiv preprint arXiv:2309.02142*.

Kearns, M., & Roth, A. (2019). The ethical algorithm: The science of socially aware algorithm design. *Oxford University Press*.

Kuhlman, C., Jackson, L., & Chunara, R. (2020). No computation without representation: Avoiding data and algorithm biases through diversity. arXiv preprint arXiv:2002.11836.

Latour, B. (2004). Why has critique run out of steam? From matters of fact to matters of concern. *Critical inquiry*, 30(2), 225-248.

Leavy, S., & O'Sullivan, B. (2020). Data, Power and Bias in Artificial Intelligence. *arXiv preprint arXiv:2008.07341*.

Li, M., Enkhtur, A., Yamamoto, B. A., & Cheng, F. (2023). Potential Societal Biases of ChatGPT in Higher Education: A Scoping Review. *arXiv preprint arXiv:2311.14381*.

Li, P., Yang, J., Wierman, A., & Ren, S. (2023). Towards Environmentally Equitable AI via Geographical Load Balancing.

Lu, G., Li, S., Mai, G., Sun, J., Zhu, D., Chai, L., ... & Liu, T. (2023). AGI for Agriculture. *arXiv preprint arXiv:2304.06136*.

Madaio, M., Blodgett, S. L., Mayfield, E., & Dixon-Román, E. (2021). Beyond "fairness": Structural (in) justice lenses on ai for education. In *The ethics of artificial intelligence in education* (pp. 203-239). Routledge.

Mak, K. K., & Pichika, M. R. (2019). Artificial intelligence in drug development: present status and future prospects. *Drug discovery today*, *24*(3), 773-780.

Mangal, A., & Kumar, N. (2016). Using big data to enhance the bosch production line performance: A kaggle challenge. In *2016 IEEE international conference on big data (big data)* (pp. 2029-2035). IEEE.

Mansbridge, J., Bohman, J., Chambers, S., Estlund, D., Føllesdal, A., Fung, A., ... & Martí, J. L. (2012). A systemic approach to deliberative democracy. *Deliberative systems: Deliberative democracy at the large scale*, 1-26.

Maple, C., Szpruch, L., Epiphaniou, G., Staykova, K., Singh, S., Penwarden, W., ... & Avramovic, P. (2023). The AI revolution: opportunities and challenges for the finance sector. *arXiv preprint arXiv:2308.16538*.

Martini, M., Cerrato, S., Salvetti, F., Angarano, S., & Chiaberge, M. (2022). Position-agnostic autonomous navigation in vineyards with deep reinforcement learning. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)* (pp. 477-484). IEEE.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, *1*(11), 501-507.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1-21.

Mittelstadt, B., & Floridi, L. (2016). The ethics of big data: current and foreseeable issues in biomedical contexts. *Science and engineering ethics*, 22(2), 303-341.

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. PublicAffairs.

Mouffe, C. (2013). Agonistics: Thinking the world politically. *Verso Books*.

Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V., ... & Zaharia, M. (2021). Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*(pp. 1-15).

Nelson, J. P., Biddle, J. B., & Shapira, P. (2023). Applications and Societal Implications of Artificial Intelligence in Manufacturing: A Systematic Review. *arXiv preprint arXiv:2308.02025*.

Niemeyer, S. (2014). Scaling up deliberation to mass publics: Harnessing mini-publics in a deliberative system. *Science Communication*, 36(4), 506-534.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Ong, K. S. H., Niyato, D., & Yuen, C. (2020). Predictive maintenance for edge-based sensor networks: A deep reinforcement learning approach. In *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)* (pp. 1-6). IEEE.

Pant, A., Hoda, R., Tantithamthavorn, C., & Turhan, B. (2022). Ethics in AI through the Developer's Prism: A Socio-Technical Grounded Theory Literature Review and Guidelines. *arXiv preprint arXiv:2206.09514*.

Park, S. (2023). Bridging the Global Divide in Ai Regulation: A Proposal for Contextual, Coherent, and Commensurable Framework.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug discovery today*, *26*(1), 80.

Przybylski, A. K., & Weinstein, N. (2013). Can you connect with me now? How the presence of mobile communication technology influences face-to-face conversation quality. *Journal of Social and Personal Relationships*, *30*(3), 237-246.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).

Rasouli, M., Chiruvolu, R., & Risheh, A. (2023). AI for Investment: A Platform Disruption. *arXiv preprint arXiv:2311.06251*.

Rawls, J. (1971). A theory of justice. *Cambridge (Mass.)*.

Resnick, M., & Wilensky, U. (1998). Diving into complexity: Developing probabilistic decentralized thinking through role-playing activities. *The Journal of the Learning Sciences*, *7*(2), 153-172.

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, 29(2), 353.

Sunstein, C. R. (2017). # Republic: Divided democracy in the age of social media. *Princeton University Press*.

Sykas, D., Sdraka, M., Zografakis, D., & Papoutsis, I. (2022). A sentinel-2 multiyear, multicountry benchmark dataset for crop classification and segmentation with deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 3323-3339.

Taiwo, E., Akinsola, A., Tella, E., Makinde, K., & Akinwande, M. (2023). A Review of the Ethics of Artificial Intelligence and its Applications in the United States. *arXiv preprint arXiv:2310.05751*.

Ting, D. S. W., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., ... & Wong, T. Y. (2018). Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*.

Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Hachette UK, 2019.

Tucker, A. D., Anderljung, M., & Dafoe, A. (2020). Social and governance implications of improved data efficiency. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 378-384).

Turkle, S. (2011). Alone together: Why we expect more from technology and less from each other. *Basic books*.

Turkle, S. (2015). Stop googling. Let's talk. *The New York Times*, *27*.

Vallor, S. (2016). Technology and the virtues: A philosophical guide to a future worth wanting. *Oxford University Press*.

Venkatasubramanian, S., Bliss, N., Nissenbaum, H., & Moses, M. (2020). Interdisciplinary Approaches to Understanding Artificial Intelligence's Impact on Society. *arXiv preprint arXiv:2012.06057*.

Viberg, O., Cukurova, M., Feldman-Maggor, Y., Alexandron, G., Shirai, S., Kanemune, S., ... & Kizilcec, R. F. (2023). Teachers' trust and perceptions of AI in education: The role of culture and AI self-efficacy in six countries. *arXiv preprint arXiv:2312.01627*.

Victor, B., He, Z., & Nibali, A. (2022). A systematic review of the use of Deep Learning in Satellite Imagery for Agriculture. *arXiv preprint arXiv:2210.01272*.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., ... & Holmberg, K. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11(1), 1-10.

Wang, S., Cooper, N., Eby, M., & Jo, E. S. (2023). From Human-Centered to Social-Centered Artificial Intelligence: Assessing ChatGPT's Impact through Disruptive Events. *arXiv preprint arXiv:2306.00227*.

Winner, L. (2010). *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press.

Woolf, B. P. (2010). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.

Yi, X., Yao, J., Wang, X., & Xie, X. (2023). Unpacking the Ethical Value Alignment in Big Models. *arXiv preprint arXiv:2310.17551*.

Young, I. M. (2000). Inclusion and democracy. *Oxford University Press*.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proceedings of the 26th International Conference on World Wide Web*, 1171-1180.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107-115.

蘇經天(Su, 2023).*新 AI 與新人類：學習、認知與生命的進化新路程*.臺灣：大塊文化.ISBN: 9786267317693.

Website & Internet Sources

[1]中研院正式開源釋出繁中優化的 Llama 2 大型語言模型,正式採用 Apache2.0 釋出 – iThome. https://www.ithome.com.tw/news/159166.Extracted Date: October 17, 2023.

[2] BNN. (2023, October 9)Academia Sinica’s Language Model Flaw: AI Responds with Unexpected Nationality.https://bnnbreaking.com/tech/ai-ml/academia-sinicas-language-model-flaw-ai-responds-with-unexpected-nationality/. Extracted Date: December 29, 2023

[3]ithome 新聞網 (2023). 中研院大型語言模型事件引發社會關注. https://www.ithome.com.tw/news/159231.Extracted Date: October 17, 2023.

[4]ithome 新聞網 (2023).明清研究專用非通用！使用開源簡中語料微調 LLM 模型引起熱議,中研院宣布已下架繁中優化的大型語言模型 CKIP-Llama-2-7b. https://www.ithome.com.tw/news/159198.Extracted Date: October 17, 2023.

[5] ithome 新聞網 (2023, October 9). 明清研究專用非通用！使用開源簡中語料微調 LLM 模型引起熱議,中研院宣布已下架繁中優化的大型語言模型 CKIP-Llama-2-7b. https://www.ithome.com.tw/news/159198

[6]ithome 新聞網 (2023). 中研院大型語言模型事件引發社會關注. https://www.ithome.com.tw/news/159231.Extracted Date: October 17, 2023.

[7] Edge AI Forum (2023).中研院繁中大語言模型引熱議 ,企業使用 LLM 該注意哪些事？ https://edge.aif.tw/ckip-llama27b-react/.Extracted Date: October 17, 2023.

[8] Edge AI Forum. (2023). 中研院繁中大語言模型引熱議 ,企業使用 LLM 該注意哪些事？https://edge.aif.tw/ckip-llama27b-react/. Extracted Date: October 17, 2023.

[9] 行政院國科會. (2019). 人工智慧科研發展指引(AI Technology R&D Guidelines).https://www.nstc.gov.tw/nstc/attachments/53491881-eb0d-443f-9169-1f434f7d33c7.Extracted Date::October 17, 2023.