computer sciences and mathematics forum

MDPI

Proceeding Paper

Towards Ethical Engineering: Artificial Intelligence as an Ethical Governance Tool for Emerging Technologies †

Dazhou Wang

School of Humanities, University of Chinese Academy of Sciences, Beijing 100049, China; dzwang@ucas.ac.cn

† Presented at Forum on Information Philosophy—The 6th International Conference of Philosophy of Information, IS4SI Summit 2023, Beijing, China, 14 August 2023.

Abstract: As a governance framework for emerging technologies, the responsible research and innovation (RRI) approach faces some fundamental conflicts, particularly between "inclusivity" and "agility". When we try to apply RRI principles to the development of AI, we also encounter similar difficulties. Therefore, it may be helpful to change the approach by not only seeing AI as the object of ethical governance but also as an effective tool for it. This involves mainly three levels: first, using AI directly to solve general ethical problems; second, using AI to solve the ethical problems brought about by AI; and third, using AI to upgrade the RRI framework. By doing something at these three levels, we can promote the fusion of AI and technology ethics and move towards ethical engineering, thus pushing ethical governance to new heights. In traditional ethical governance approaches, ethics is external, brought in by external actors, and the focus is on actors. However, in this new approach, ethics must be internalized in technology, with the focus not only on actors, but also on technology itself. Its essence lies in the invention and creation of technologically ethical governance tools.

Keywords: RRI; artificial intelligence; ethical governance; ethical engineering

1. Introduction

Due to their novel, variable, and uncertain development prospects, emerging technologies cannot have clear measurement standards at the outset. Therefore, innovators face inescapable uncertainty and can only move forward. Once emerging technologies are introduced, they will nonlinearly interact with various economic and social elements, constantly undergoing "translation" by other actors, inevitably leading to unforeseeable economic and social consequences, including uncontrollable risks and resulting in governance dilemmas that are difficult to solve [1]. People either allow these emerging technologies to develop freely and only address negative consequences after they occur, in which case the harm has already been carried out, or strictly regulate them from the outset, which might stifle their development—this is known as the Collingridge dilemma [2]. There have been many studies on how to overcome this dilemma, and "responsible research and innovation" (RRI) has become the most popular approach in recent years.

As a new concept that has emerged in recent years in European and American countries, RRI requires attention to social, ethical, and legal issues in the research and innovation process; emphasizes the participation and collective negotiation of stakeholders as early as possible; and encourages the early intervention and real-time evaluation of humanists such as ethicists [3]. However, the RRI framework faces some problems, and the most serious one is that there is an inherent contradiction between the inclusiveness of stakeholders and the agility of collective action. This framework requires as many stakeholders as possible to participate. Due to the different interests and values of each stakeholder, there will be many internal conflicts, which will cause the collective action to be slow and make it difficult to conduct innovative activity effectively. That is, it is difficult to reach a consensus and difficult to take action, so many innovations may not be able to be

Citation: Wang, D. Towards Ethical Engineering: Artificial Intelligence as an Ethical Governance Tool for Emerging Technologies. *Comput. Sci. Math. Forum* **2023**, *8*, 76. https://doi.org/10.3390/cmsf2023008076

Academic Editors: Zhongzhi Shi and Wolfgang Hofkirchner

Published: 10 October 2023

carried out if the RRI principle is completely implemented. Therefore, simply internalizing outsiders and turning them into insiders to some extent is not enough to fundamentally solve the problem. As we aim to implement the RRI concept in the development of artificial intelligence (AI), we inevitably encounter similar challenges [4], namely, that recruiting external parties alone cannot guarantee responsible AI development. It is crucial to address this issue meaningfully.

2. Turning Artificial Intelligence into a Tool for Ethical Governance

To address this problem, we should shift our thinking mode. Rather than treating artificial intelligence solely as a subject of ethical governance, we should explore how we can turn it into a tool for ethical governance, thereby promoting RRI frameworks to be upgraded. This consists of three levels. Firstly, we could use artificial intelligence directly to solve general ethical problems. Secondly, we could use artificial intelligence to address the ethical problems that it itself has caused. Thirdly, we could use artificial intelligence to enhance RRI frameworks. By taking action on these three levels, we may be able to develop a new field of ethical engineering.

2.1. Using AI to Solve General Ethical Issues in Science and Technology

The first step is to use artificial intelligence to solve general ethical problems in science and technology. For example, there are lots of academic misconduct cases in the research field, including plagiarism. It used to be difficult to detect plagiarism, but now things are different. There are plagiarism detection tools available, and in general, the first step after a journal receives a paper submission is to check for duplicates. However, this retrieval system still has problems. Basically, it can only detect textual duplications and not semantic duplications. In other words, the expression can be changed and passed, but the idea itself may still be plagiarized. Moreover, most plagiarism detection tools can only work for documents in a certain language and cannot detect plagiarism with cross-language retrieval. As a result, there may be a situation where published articles in a foreign language are copied and assembled in the home country, or overseas students translate publications in their native language into foreign languages to apply for their degrees. Through natural language processing technologies, especially large language models (LLM) [5], new technologically ethical tools can be developed to achieve cross-language retrieval and semantic duplication detection. With this, academic misconduct will be difficult to hide.

Actually, an ethical AI machine can function as a universal tool for ethical discussions with researchers and innovators, a conversational partner in people's daily ethical decision-making processes. Current efforts are underway to develop a range of digital thinkers, including digital philosophers, ethicists, and even common citizens. By engaging in real-time conversations, these machines can prompt researchers, innovators, and the public to consider ethical issues and remain vigilant about ethical considerations. Individuals can seek guidance from an AI-based tool like ChatGPT or digital ethicists whenever confronted with ethical issues. This ethical AI has the potential to expand its reach throughout society and provide support in addressing a multitude of ethical concerns.

2.2. Using AI to Solve Ethical Issues Caused by AI

The second aspect is to use AI to solve ethical issues related to Artificial Intelligence. The basic idea is to use AI to counter AI. When we try to govern AI, we should not forget that AI is also a tool for governing AI. This actually represents a kind of reflexivity—AI is used to control AI, and by doing so, a cycle is formed where AI develops along an ethical development trajectory through iterative learning. For instance, we can develop an AI ethics verification machine to check AI ethics during the AI development process. In general, the ethical principles can be converted into industry standards on which the ethical monitoring machine can be built. Such an ethical machine can penetrate all aspects of AI applications so as to serve as a supporting platform, which could be used by AI developers,

customers, and even third parties. In this way, AI ethics can be transformed into a concrete technical tool. This kind of technical tool is a distributed platform, a real-time operational tool, and can be understood as a kind of ethical machine.

Combating AI is a complex task that requires consideration of multiple factors such as data security, network security, and algorithm security. The overarching method for leveraging AI in this pursuit involves: Using AI algorithms to identify abnormal activities such as threats, attacks, and malicious programs that steal data from the network; Using AI to simulate attacks so potential vulnerabilities and security risks of AI-based systems can be identified and corresponding preventive measures can be taken; Using AI to establish a secure network area to prevent unauthorized access and other threats. In short, to address the ongoing threats posed by AI, creative and innovative AI-based strategies should be continuously developed.

2.3. Using AI to Upgrade the RRI Framework

The third point is to upgrade the RRI framework by using artificial intelligence. Utilizing AI can help alleviate the contradiction between inclusiveness and agility, promoting both inclusiveness and agility simultaneously. According to the AI framework, many stakeholders should be included, but there is a coordination problem after they are brought in. If coordination is only carried out in person, the cost will be very high. If a network platform based on AI is established, an auditable stakeholder network can be formed, and coordination among stakeholders can be more convenient and efficient. This approach can not only include more stakeholders but also greatly reduce coordination costs, thus simultaneously enhancing the inclusiveness and agility of responsible innovation.

As we know, typical RRI procedures include four dimensions: anticipation, reflexivity, engagement, and feedback. The anticipation dimension requires the combination of scientific evidence and future analysis to enable innovators to better understand the opportunities and challenges they face; the reflexivity dimension requires innovators to reflect on their own behavior and innovation process; the engagement dimension requires placing visions, purposes, problems, and difficulties in a larger social context and achieving collective deliberation through participation; and the feedback dimension refers to adjusting the framework and direction of innovation activities in a timely manner based on voices from stakeholders [6]. When discussing these four dimensions in the past, people never considered the issues of "cost" and "feasibility" of implementation. Based on artificial intelligence, all four dimensions can be upgraded to become AI-based anticipation, AI-based reflexivity, AI-based engagement, and AI-based feedback. AI tools can be applied in each dimension, allowing for digital philosophers, digital ethicists, digital sociologists, digital lawyers, and even digital citizens to participate. With the increasing use of digital stakeholders, the process of communication and dialogue is moving towards real-time automation, which can significantly reduce operational costs. Nonetheless, it is not an indicator that real actors' involvement is dispensable; rather, it is the first step in implementing RRI. This approach effectively addresses the conflict between inclusivity and agility and ensures responsible innovation is not merely rhetoric but is practiced in reality.

3. Towards Ethical Engineering

By utilizing a positive approach, the field of technology ethics can be transformed from a restrictive one to an empowering one. In the past, ethical discussions surrounding technologies were typically focused on limiting the choices of innovators, dictating what they should not do, or restricting the potential development of technologies. This approach was overly negative and constrictive. However, technology ethics can also serve as a positive and inspirational tool, motivating innovators to take affirmative steps toward realizing the ethical potential of technology and innovation [7].

In fact, technology itself has ethical potential, and it plays a crucial role in ethical governance. Discussions on the ethical governance of AI also need to be elevated to this level in order to fully utilize the ethical potential of AI. As an ethical imperative, the

development and application of AI should stimulate the potential for human development and the ethical potential of AI technology. For example, following the emergence of AlphaGo, even top Go masters were unable to defeat it. Yet people's enthusiasm for playing Go has not diminished. Instead, players have turned to following AlphaGo's lead to develop new techniques. Similarly, the currently popular ChatGPT is about to become everyone's assistant, conversation partner, and even tutor. Such a kind of AI machine naturally requires the injection of knowledge, emotion, and even morality, which inevitably entails some risks [8]. However, from the perspective of iterative learning, ethical governance tools can be gradually developed to handle such risks as they emerge throughout the AI development process instead of expecting to solve all problems at the outset [9].

To achieve this, the focus will be shifted from technology ethics to ethical technology. Rather than focusing primarily on innovators, as technology ethics requires, we should turn to the development of ethical technologies, which involves technology-based ethics or incorporating ethics into technology design. This approach views AI as an opportunity to proactively address ethical concerns and emphasizes the fusion of AI and technology ethics. While traditional considerations view ethics as an external element that is brought in by external actors, with the focus centered on innovators [10], the new perspective emphasizes the need to internalize ethics in technology, with a focus not only on innovators but also on technology itself. This requires the development of ethical governance tools through the invention and creation of technology-based ethical solutions. It is an engineering challenge to effectively integrate ethical considerations into technology design, making it an essential aspect of the governance of technological developments.

The establishment of ethical governance through engineering is poised to emerge as a new direction in engineering sciences, namely ethical engineering. The need for ethics-based AI is paralleled by the need for AI-based ethics. With this in mind, an AI-based ethical governance platform can be developed, encompassing various industries and society at large. To achieve this, a new interdisciplinary research field combining ethics and technology should be established, led by a combination of ethicists and engineering scientists. Both researchers and research management departments bear the ethical responsibility of collaborating with ethicists to develop such a new discipline so as to collectively create an ethical platform applicable to society as a whole.

4. Conclusions

In conclusion, the solution to resolving ethical concerns associated with technology, such as those related to artificial intelligence, lies in using artificial intelligence to create governance tools that prioritize ethics, specifically AI-based ethics tools. This represents a significant ethical responsibility for researchers, engineering professionals, and management teams at various levels.

Although science and technology management has made significant contributions to promoting ethics by establishing ethical guidelines and monitoring the ethical behaviors of scientists and innovators, it is not enough to stop at this level. AI should be regarded as an opportunity to proactively address ethical concerns and aggressively pursue research on AI-based ethics. This endeavor will require substantial support to materialize.

Funding: This research was funded by the National Social Science Foundation of China (No. 19ZDA040).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study did not create new data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kuhlmann, S.; Stegmaier, P.; Konrad, K. The tentative governance of emerging science and technology—A conceptual introduction. *Res. Policy* **2019**, *48*, 1091–1097. [CrossRef]
2. Collingridge, D. *The Social Control of Technology*; Pinter: London, UK, 1980.
3. Gardner, J.; Williams, C. Responsible research and innovation: A manifesto for empirical ethics. *Clin. Ethics* **2015**, *10*, 5–12. [CrossRef]
4. Chen, L.; Wang, D. Research Progress of Brain-Computer Interface Responsible Innovation. *J. Eng. Stud.* **2019**, *11*, 390–399. [CrossRef]
5. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2023**, arXiv:2303.18223.
6. Stilgoe, J.; Owen, R.; Macnaghten, P. Developing a framework for responsible innovation. *Res. Policy* **2013**, *42*, 1568–1580. [CrossRef]
7. Wang, D. Towards Responsible Engineering: Interpretation and Implementation of Ethical Codes. *Chem. Eng. Educ.* **2020**, *37*, 1–7.
8. Zhao, T. How is artificial intelligence's self-awareness possible? *J. Dialectics Nat.* **2019**, *41*, 1–8.
9. Wang, D. Toward an Experimental Philosophy of Engineering. In *Philosophy of Engineering: East and West*; Springer: Berlin/Heidelberg, Germany, 2018.
10. Chen, X. *Ethics Guide for Artificial Intelligence*; University of Science and Technology of China Press: Hefei, China, 2021.