computer sciences and mathematics forum

MDPI

Proceeding Paper

Ethical Governance of AI: An Integrated Approach via Human-in-the-Loop Machine Learning †

Ximeng Chen

Department of Philosophy, Xi'an Jiaotong University, Xi'an 710049, China; chenximeng@xjtu.edu.cn

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

Abstract: AI technology is empowering human society. As for the ethical issues arising from technology, there have been many concerns and discussions, and the corresponding governance approaches have been put forward. However, there is not yet an integrated approach. The known governance approaches have a variety of flaws, leading to theoretical problems such as existence, interpretation, and value, and practical problems such as justice, security, and responsibilities. To build an integrated AI governance approach considering multi-stakeholder, an ethical governance approach of AI named "human-in-the-loop" is proposed in this paper. It analyzes the advantages of this method and points out future problems.

Keywords: artificial intelligence; human-in-the-loop; ethical governance

1. Introduction

AI science and technology enable human beings to manipulate a great many aspects of life effectively [1–6]. With the generation of ChatGPT and other generative AI [7], big data-driven knowledge learning [8], cross-media collaborative processing [9], augmented intelligence [10], collective intelligence, and autonomous system [11] have attracted so much attention. The new generation of AI plays multiple critical roles in our daily life [1–11]. Correspondingly, the diversity and uncertainty of future AI development will bring new challenges to its ethical governance as well, and it is urgent to ensure the credible, controllable, and sustainable development of AI [12–19].

To create an AI that is more responsible than current, achieving effective ethical embedding should be the primary mission [20]. In the technical process, the key step of ethical embedding is data annotation. There are two approaches, one is the top-down human annotation (expert system of symbolist approaches, such as Bacon, MedEthEx, etc.), and the other is the bottom-up machine annotation (such as Driverless vehicles, Ask Delphi, etc.) [14], both of them have their disadvantages. The former has high requirements for the annotation, but the annotation process will inevitably be influenced by the personal will, moral cultivation, and values of the annotation. If the annotation is not ethical enough, then the machine designed will not be ethical. This will affect the fairness and consistency of the system. The latter is sensitive to the data learned by the algorithm, but the machine learning process is affected by multiple factors such as human behavior, semantic recognition, algorithm optimization approach, and the environment, which affects the security and controllability of AI. The inappropriate data annotation approach will lead to the lack of accuracy, robustness, and comprehensiveness of AI in ethical judgment and decision-making. This will create unethical AI products. To improve that situation, this paper put forward a human-in-the-loop AI ethics embedded method, and analyzes its characteristics and feasibility. Finally, it explains why this method can be used as a new type of AI governance approach and play an important role in the future.

check for updates

Citation: Chen, X. Ethical Governance of AI: An Integrated Approach via Human-in-the-Loop Machine Learning. *Comput. Sci. Math. Forum* **2023**, *8*, 29. https://doi.org/10.3390/cmsf2023008029

Academic Editors: Zhongzhi Shi and Wolfgang Hofkirchner

Published: 11 August 2023

Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

2. Problems

Because AI can't make a moral judgment, it is urgent to improve its autonomy in self-recognition and scene recognition. This means breaking through the technical bottleneck of machine learning and improving its transparency and security performance. Both human annotation and machine annotation have possibilities to embed ethics and achieve "good", but they also have corresponding flaws, which makes ethics embedding difficult. On one hand, we need the transparency and interpretability brought by human annotation to expect AI to operate stably and continuously. On the other hand, we also need the flexibility and efficiency brought by machine annotation to adapt to the complex ethical context. In one short, we need to thoroughly combine human intelligence with machine intelligence.

At present, although AI has been proven to achieve the enhancement and expansion of human ability to a certain extent, humans know little about the various possibilities of machine intelligence. The different values and demands of different stakeholders also mean that a universal ethical framework is difficult to establish, and the development of this technology is increasingly out of control. Therefore, as active moral agents, human beings should assume the responsibility of prediction, design, and supervision for the development and management of AI technology. It is necessary to explore the appropriate AI ethical embedding method (annotation method) and governance approach, to promote the integration of human-machine value and establish a comprehensive framework for AI ethical governance (Figure 1).



Figure 1. The stakeholders in AI ethical governance.

3. Solution

Human-in-the-loop is a semi-supervised human-computer interaction learning method [21], which aims at achieving the accuracy of machine learning and assisting human learning. It implements the high efficiency and accuracy of AI in decision-making through scene recognition, high-quality human annotation, and active learning. A possible ethical solution can be provided by using the steps as follows (Figure 2):

- Scene recognition: For a certain scene of AI decision-making, try to design an online questionnaire, and open it to the public to collect their views;
- Marks: Mark the original data of step 1, use machine learning to mark some of the processes of semi-automation, or help to improve the efficiency. In the context of ethical decisions, the best practice is to implement the " loop" and annotate the data correctly, using the annotated data to train the model, then use the trained model to sample more data for labeling;
- Active learning: Combining the sampling strategy of diversity, uncertainty, and randomness, the results of the online questionnaire in step 1 were sampled and analyzed, and adjusted to the sampling strategy of active learning. The iterative process in step 2

is repeated until an artificial decision-making process that is most close to a real-world scenario is generated and ethically justified.

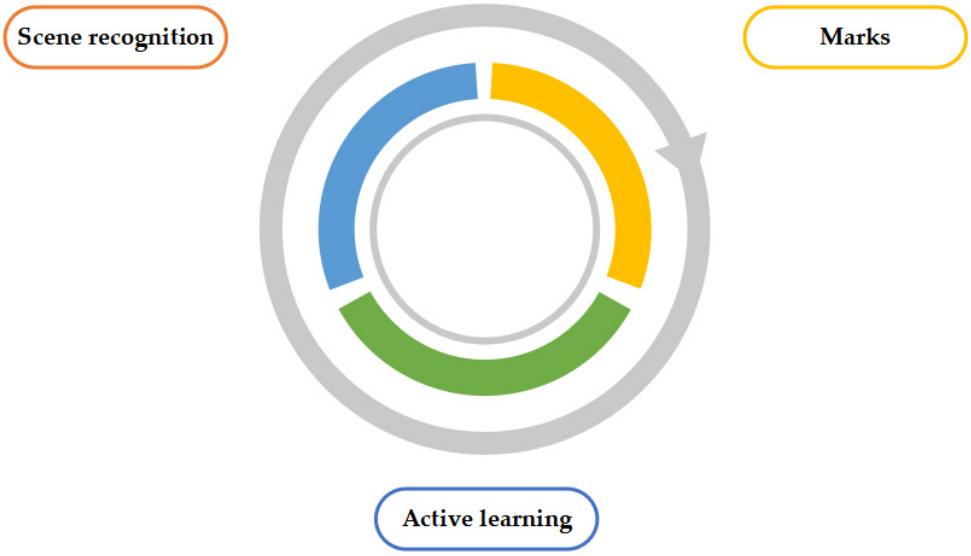


Figure 2. The procedure of Human-in-the-Loop.

Funding: This research was funded by the Soft Science Research Plan of Shaanxi Province, grant number 2023-CX-RKX-140, and the 2022 Basic Scientific Research Project of Xi'an Jiaotong University, grant number SK2022047.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Taddeo, M. The ethical governance of the digital during and after the COVID-19. *Minds Mach.* **2020**, *30*, 171–176. [CrossRef] [PubMed]
2. Hilb, M. Toward artificial governance? The role of artificial intelligence in shaping the future of corporate governance. *J. Manag. Gov.* **2020**, *24*, 851–870. [CrossRef]
3. Floridi, L. Artificial intelligence as a public service: Learning from Amsterdam and Helsinki. *Philos. Technol.* **2020**, *33*, 541–546. [CrossRef]
4. Munoko, I.; Brown-Liburd, H.L.; Vasarhelyi, M. The ethical implications of using artificial intelligence in auditing. *J. Bus. Ethics* **2020**, *167*, 209–234. [CrossRef]
5. Currie, G.; Hawk, K.E.; Rohren, E.M. Ethical principles for the application of artificial intelligence (AI) in nuclear medicine. *Eur. J. Nucl. Med. Mol. Imaging* **2020**, *47*, 748–752. [CrossRef] [PubMed]
6. Buller, T. Brain-Computer Interfaces and the Translation of Thought into Action. *Neuroethics* **2021**, *14*, 155–165. [CrossRef]
7. Weisz, J.D.; Muller, M.; He, J.; Houde, S. Toward General Design Principles for Generative AI Applications. *arXiv* **2023**, arXiv:2301.05578.
8. Adam, K.; Bakar, N.A.; Fakhreldin, M.A.; Majid, M.A. Big Data and Learning Analytics: A Big Potential to Improve E-Learning. *J. Comput. Theor. Nanosci.* **2018**, *24*, 7838–7843. [CrossRef]
9. Fang, M. Intelligent Processing Technology of Cross Media Intelligence Based on Deep Cognitive Neural Network and Big Data. In Proceedings of the 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 23–25 October 2020. [CrossRef]
10. Kim, J.; Davis, T.; Hong, L. Augmented Intelligence: Enhancing Human Decision Making. In *Bridging Human Intelligence and Artificial Intelligence. Educational Communications and Technology: Issues and Innovations*; Albert, M.V., Lin, L., Spector, M.J., Dunn, L.S., Eds.; Springer: Cham, Switzerland, 2022. [CrossRef]
11. Malone, T.W.; Bernstein, M.S. *Handbook of Collective Intelligence*; MIT Press: Cambridge, MA, USA, 2015; pp. 58–87.

12. ÓhÉigeartaigh, S.S.; Whittlestone, J.; Liu, Y.; Zeng, Y.; Liu, Z. Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance. *Philos. Technol.* **2020**, *33*, 571–593. [CrossRef]
13. Baker-Brunnbauer, J. Management perspective of ethics in artificial intelligence. *AI Ethics* **2021**, *1*, 173–181. [CrossRef]
14. Wallach, W.; Allen, C. *Moral Machines: Teaching Robots Right from Wrong*; Oxford University Press: Oxford, UK; Peking University Press: Beijing, China, 2017; pp. 110–190.
15. Dignum, V. Ethics in artificial intelligence: Introduction to the special issue. *Ethics Inf. Technol.* **2018**, *20*, 1–3. [CrossRef]
16. Buruk, B.; Ekmekci, P.E.; Arda, B. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Med. Health Care Philos.* **2020**, *23*, 387–399. [CrossRef] [PubMed]
17. Cai, H. Reaching consensus with human beings through block chain as an ethical rule of strong artificial intelligence. *AI Ethics* **2021**, *1*, 55–59. [CrossRef]
18. Cath, C.; Wachter, S.; Mittelstadt, B.; Taddeo, M.; Floridi, L. Artificial intelligence and the 'Good Society': The US, EU, and UK approach. *Sci. Eng. Ethics* **2018**, *24*, 505–528. [PubMed]
19. Albrecht, J.; Kitanidis, E.; Fetterman, A.J. Despite "super-human" performance, current LLMs are unsuited for decisions about ethics and safety. *arXiv* **2022**, arXiv:2212.06295.
20. Liu, J. Human-in-the-Loop Ethical AI for Care Robots and Confucian Virtue Ethics. In Proceedings of the 14th International Conference on Social Robotics, Florence, Italy, 13–16 December 2022.
21. Monarch, R.M. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*; Manning Publications Co. LLC: Dublin, Ireland, 2021; pp. 31–63.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.