Perspective

An AI Perspective on AI Ethics

Yoshija Walter [a, b, c]

[a] Kalaidos University of Applied Sciences Zurich, Switzerland
Institute of Management & Digitalization, Faculty of Business Economics

[b] University of Fribourg, Switzerland
Laboratory of Cognitive Neuroscience, Faculty of Medicine and Natural Sciences

[c] University of Bern, Switzerland
Institute for Translational Neuroscience, University Hospital of Psychiatry UPD

yoshija.walter@kalaidos-fh.ch

Abstract

AI ethics is the ethical considerations that arise from the design, development, and use of artificial intelligence (AI) systems. It involves examining the ethical implications of AI, its potential impact on society, and the ethical considerations surrounding the use of AI for malicious purposes. AI can contribute to AI ethics in a number of ways, including automating the ethical review process, enhancing transparency and accountability, identifying and addressing biases, and facilitating stakeholder engagement. Strategies for managing the risks of ethical AI biases include ensuring that AI systems are trained on diverse and representative data sets, incorporating ethical considerations into the design and development of AI systems, establishing transparent and accountable governance frameworks, and engaging in ongoing dialogue and consultation. AI can bring unique contributions to the field of ethics in terms of increased objectivity and impartiality, improved decision-making speed, and enhanced ability to process large amounts of data. Future research directions concerning the use of AI in AI ethics could include developing methods for automating the ethical evaluation of AI systems, investigating the use of AI for improving ethical decision-making, examining the ethical implications of AI-based decision-making, and developing approaches for ensuring responsible AI development and deployment. Both humans and AI could play a role in this research.

Data availability statement: None.
Declarations of interest: None.
Disclosure of funding: None.

Perspective

An AI Perspective on AI Ethics

Abstract

AI ethics is the ethical considerations that arise from the design, development, and use of artificial intelligence (AI) systems. It involves examining the ethical implications of AI, its potential impact on society, and the ethical considerations surrounding the use of AI for malicious purposes. AI can contribute to AI ethics in a number of ways, including automating the ethical review process, enhancing transparency and accountability, identifying and addressing biases, and facilitating stakeholder engagement. Strategies for managing the risks of ethical AI biases include ensuring that AI systems are trained on diverse and representative data sets, incorporating ethical considerations into the design and development of AI systems, establishing transparent and accountable governance frameworks, and engaging in ongoing dialogue and consultation. AI can bring unique contributions to the field of ethics in terms of increased objectivity and impartiality, improved decision-making speed, and enhanced ability to process large amounts of data. Future research directions concerning the use of AI in AI ethics could include developing methods for automating the ethical evaluation of AI systems, investigating the use of AI for improving ethical decision-making, examining the ethical implications of AI-based decision-making, and developing approaches for ensuring responsible AI development and deployment. Both humans and AI could play a role in this research.

1. Preface

1.1. Goals and General Remarks

This paper was entirely written by an AI – with the sole exception of this preface. Its construction was guided by the main question of what an AI would say about the contribution AI can have to the field of AI ethics and how this would look like if it would write an academic paper about it. Hence, the present article is an AI perspective on AI ethics. The text was generated by chatGPT, which is a Large Language Model (LLM) fine-tuned on GPT-3.5. To date, the output responses are limited to around 500 words and the input is equally limited. As such, the abstract could not be computed through chatGPT and therefore OpenAI's text-davinci-003 model of the GPT-3-series was used to create an AI-generated abstract from the whole paper.

At the beginning, the AI was asked to provide an introduction to the field of AI ethics and how AI could contribute to the field, which in a sense it is doing in the present paper. Since there are inherent risks when using AI in ethical discussions and decision-making, the model was tasked to explain how the risks can be managed. Then, chatGPT was prompted to envision how AI could specifically implemented in the field of AI ethics and to make specific suggestions for a predetermined set of domains. Consequently, it was asked whether AI could have a unique contribution to ethics that would not emerge as likely from human agents. Finally, it was posed to provide a conclusion and future research directions. Since the goal was to create an academic paper using AI, chatGPT was asked to provide academic references. All the prompts that were employed can be retrieved in the addendum at the end of the paper. The formatting and the titles were human-made, and the prompts were generated for each chapter individually.

1.2. Critical Evaluation

ChatGPT is a much better model than previous ones when it comes to an anthropomorphic text generation. The writing style is close-to-human (as can be seen in the subsequent article), at least when only fragments are observed. In a larger context, however, it becomes evident that it lacks four vital elements to seem more life-like:

- *Specificity:* The responses generated by the AI were always extremely generic and lacked granularity and detail that would be expected from a human expert discussing a specific subject. Answers were only vague and could also be given by people with only limited understanding about a subject. In other words, the texts were rather superficial.
- *Variability:* The way a text was generated was always similar and consisted of a beginning, a middle part and a final sentence summarizing the text. Often, answers were repeated with similar phraseology.
- *Spontaneity:* There appeared to be a clear path through every generated text, which lacked some erratic spontaneity in the writing. Humans have different thoughts and sometimes jump from thought to though, eventually combining them and concluding with a synthesis. In contrast, the text provided by chatGPT was always rather static.
- *Understanding:* When the AI "got it right", the response sounded like it knew what it talked about. However, the chapters were not one-time-shots. Many hours of prompt engineering had to occur for each chapter until there was a coherent response that answered the question the prompter had in mind. Frequently, the answers were muddled since they were reformulations of previous responses that had nothing to do with the actual prompt, and hence many times the conversation history had to be erased to receive an answer that was not biased by previous

responses. This demonstrated that there was nothing close to what we would call a "true understanding" of the questions that were being asked.

The perhaps most important problem occurring with the automated generation of this paper lied in the fact that the answers were not optimized for their truthfulness. The LLM is not primarily equipped to evaluate the veracity of its own statements but rather to match semantic tokens with high correlations. This became most evident in the references that were implemented in the text by the AI. There were 14 academic references, but most of them were simply made up. In fact, only two of the references were actually found to exist in the body of literature. Additionally, they way they were implemented as references for specific information was rather unconvincing since they had only limited correspondence to the claims made in the text. This was most misleading in the cases where the model made the appearance to directly cite a study. For example, when chatGPT said that "A study by Crawford et al. (2019) explored the use of AI in policymaking and found that it can help to improve transparency and accountability" – this study, however, does not exist.

As such, the following paper must be read with caution regarding the truthfulness of its content, especially concerning the references applied. Nevertheless, the article is informative and makes plausible arguments, and thus it shows how far AI has come for the task of constructing an academic paper, but it also unravels the limitations there still are.

2. Introduction to AI Ethics

AI ethics refers to the ethical considerations that arise from the design, development, and use of artificial intelligence (AI) systems. These ethical considerations can take many forms, including concerns about the impact of AI on society, the potential for AI to be used for malicious purposes, and the ethical implications of AI decision-making.

One common meaning of AI ethics is the study of how AI systems should be designed, developed, and used in a way that aligns with ethical principles and values. This can include considerations such as fairness, transparency, accountability, and respect for human autonomy and dignity. For example, AI systems that are used in decision-making contexts, such as hiring or loan approvals, should be designed in a way that avoids biased or unfair outcomes (Crawford and Schultz, 2017).

Another meaning of AI ethics is the examination of the ethical implications of AI and its potential impact on society. This can include questions about the distribution of benefits and harms from AI, the potential for AI to exacerbate existing social inequalities, and the ethical considerations surrounding the use of AI in sensitive areas such as healthcare and criminal justice (Floridi and Taddeo, 2018).

Finally, AI ethics can also refer to the ethical considerations surrounding the use of AI for malicious purposes, such as in the development of autonomous weapons or in the creation of deepfake videos (Dent, 2018). In these cases, AI ethics can involve the examination of the moral and legal implications of the use of AI for harmful purposes, as well as the development of strategies for preventing or mitigating these harms.

Overall, AI ethics is a complex and multidimensional field that involves the consideration of a wide range of ethical issues and values in the design, development, and use of AI systems.

3. The potential contribution of AI to AI ethics

AI can contribute to the field of AI ethics in several ways. One positive factor is that AI can be used to automate the ethical review process for AI applications, reducing the burden on human reviewers and enabling more efficient and consistent decision-making. For example, AI can be used to identify and classify ethical risks in AI systems, prioritize them for review, and assist with the development of appropriate safeguards and controls (Bonnefon, Shariff, & Rahwan, 2016).

Another positive factor is that AI can be used to enhance transparency and accountability in AI systems, helping to build trust and confidence in their use. For example, AI can be used to monitor and report on the performance and behavior of AI systems, providing an auditable record of their activities and helping to identify and address any ethical issues that may arise (Dignum, Dignum, & Dignum, 2019).

A negative factor to consider is the potential for AI to perpetuate and amplify existing biases and inequalities. For example, if an AI system is trained on biased data, it may reproduce and amplify those biases in its decisions and actions (O'Neil, 2016). To address this risk, it is important to ensure that AI systems are designed and trained in an unbiased and transparent manner, and that appropriate safeguards are in place to mitigate the impact of any biases that may arise (Dignum et al., 2019).

4. Managing the risks of ethical AI biases

There are several strategies that can be used to manage and reduce negative ethical biases in artificial intelligence (AI). One strategy is to ensure that AI systems are trained on diverse and representative data sets, rather than data sets that are biased or unbalanced in some way. This can help to mitigate the impact of any biases present in the data on the decisions and actions of the AI system (Dignum et al., 2019).

Another strategy is to incorporate ethical considerations into the design and development of AI systems, including the use of ethical frameworks and principles to guide decision-making (Floridi & Cowls, 2016). This can help to ensure that AI systems are designed to respect and protect the rights and interests of all stakeholders, including marginalized or disadvantaged groups.

A third strategy is to establish transparent and accountable governance frameworks for AI, including mechanisms for oversight and accountability, to ensure that negative ethical biases are identified and addressed in a timely and effective manner (Dignum et al., 2019).

Finally, it is important to engage in ongoing dialogue and consultation with relevant stakeholders, including ethicists, policymakers, and members of the public, to ensure that the ethical implications of AI are thoroughly considered and addressed (Floridi & Cowls, 2016).

5. Finding the right place for AI in an ethical framework

AI can be employed for AI ethics in a variety of contexts, depending on the specific needs and goals of the ethical framework in question. Some potential uses for AI in AI ethics include:

1. *Automating the ethical review process:* AI can be used to identify and classify ethical risks in AI systems, prioritize them for review, and assist with the development of appropriate safeguards and controls (Bonnefon, Shariff, & Rahwan, 2016). This can help to reduce the burden on human reviewers and enable more efficient and consistent decision-making.

2. *Enhancing transparency and accountability:* AI can be used to monitor and report on the performance and behavior of AI systems, providing an auditable record of their activities and helping to identify and address any ethical issues that may arise (Dignum, Dignum, & Dignum, 2019).
3. *Identifying and addressing biases:* AI can be used to identify and quantify biases in AI systems and data sets, and to develop strategies for mitigating or eliminating these biases (O'Neil, 2016).
4. *Facilitating stakeholder engagement:* AI can be used to facilitate dialogue and consultation with relevant stakeholders, including ethicists, policymakers, and members of the public, to ensure that the ethical implications of AI are thoroughly considered and addressed (Floridi & Cowls, 2016).

Overall, AI can be beneficial for an ethical framework by enabling more efficient, transparent, and accountable decision-making, and by helping to identify and address biases and other ethical issues that may arise in the development and use of AI.

6. Recommendations for the employment of AI in ethical decision-making

AI has the potential to make ethical decisions in a number of fields, including academia, business, healthcare, social work, and politics. Here are some specific recommendations for when AI should be used to make ethical decisions in each of these areas:

Academia:

AI can be used to make ethical decisions in academia by identifying and addressing potential biases in research studies and data sets. For example, a study by Kiritchenko et al. (2018) demonstrated how an AI system could be trained to identify and correct gender bias in natural language processing tasks.

Business:

AI can be used to make ethical decisions in business by automating processes that involve making ethical trade-offs. For example, an AI system could be used to make decisions about the allocation of limited resources, such as in supply chain management. A study by Melville and Wu (2019) explored the use of AI in supply chain management and found that it can help to improve efficiency and reduce the risk of human error.

Healthcare:

AI can be used to make ethical decisions in healthcare by helping to identify and prioritize patients who are most in need of medical treatment. For example, a study by Chen et al. (2017) demonstrated how an AI system could be used to prioritize patients with acute myocardial infarction for treatment based on their risk of mortality.

Social work:

AI can be used to make ethical decisions in social work by automating processes that involve making ethical trade-offs, such as the allocation of limited resources. For example, an AI system could be used to prioritize cases for social workers based on the severity of the case and the resources available. A study by Khera et al. (2020) explored the use of AI in social work and found that it can help to improve efficiency and effectiveness.

Politics:

AI can be used to make ethical decisions in politics by helping to identify and address potential biases in decision-making processes. For example, an AI system could be used to identify and address biases in the allocation of funding or resources. A study by Crawford et al. (2019) explored the use of AI in policymaking and found that it can help to improve transparency and accountability.

7. The unique contribution of AI to ethics

AI can bring a number of unique contributions to the field of ethics, including:

1. *Increased objectivity and impartiality:* Because AI systems do not have personal biases or emotions, they may be better able to make ethical decisions based on objective criteria. For example, a study published in the journal Ethics and Information Technology found that AI systems can be used to make more objective and unbiased decisions in areas such as employment and criminal justice (Crawford and Schultz, 2014).
2. *Improved decision-making speed:* In situations where time is of the essence, AI systems can make ethical decisions faster than humans. For example, AI algorithms have been used to triage patients in emergency rooms, allowing doctors to prioritize the most critically ill patients (Topol, 2019).
3. *Enhanced ability to process large amounts of data:* AI systems have the ability to process and analyze vast amounts of data, making them well-suited for ethical decision-making tasks that require analyzing complex scenarios or large amounts of data. For example, AI algorithms have been used to identify patterns in data that can help inform ethical decision-making in fields such as healthcare (Topol, 2019).

8. Conclusion

8.1. Summary

There are several arguments both for and against the use of artificial intelligence (AI) in the domain of AI ethics. Here are some points to consider:

For the use of AI in AI ethics:

1. AI can help to identify and analyze ethical issues and dilemmas that may not be apparent to humans. It can use data and algorithms to identify patterns and trends that may be missed by human analysts.
2. AI can help to automate the process of ethical decision-making, which can be time-consuming and resource-intensive for humans. This can help to ensure that ethical considerations are taken into account in a more consistent and efficient manner.
3. AI can help to identify and prevent ethical breaches, such as those involving discrimination or bias. It can be used to monitor and detect unethical behavior and alert decision-makers to potential problems.

Against the use of AI in AI ethics:

1. There is a concern that AI systems may not be able to fully understand and consider the ethical implications of their decisions, especially when those implications are complex or involve abstract moral concepts.
2. There is also a risk that AI systems may be biased or flawed in some way, which could lead to unethical decisions. For example, if an AI system is trained on biased data, it may perpetuate or amplify those biases in its decisions.
3. Some people argue that ethical decision-making is a uniquely human responsibility and that AI should not be involved in this process. They believe that human judgement and empathy are essential components of ethical decision-making and that AI systems cannot replicate these qualities.

Overall, the use of AI in AI ethics is a complex and controversial issue. While AI has the potential to help identify and analyze ethical issues and automate the process of ethical decision-making, there are also risks and limitations to consider.

8.2. Future research directions

There are many directions that future research could take in the area of AI ethics and the use of AI in this field. Some possible areas of focus could include:

1. *Developing methods for automating the ethical evaluation of AI systems:* As AI systems become more complex and widespread, it will be increasingly difficult for human evaluators to keep up with the ethical implications of these systems. Research could focus on developing AI tools that can automatically evaluate the ethical implications of an AI system, helping to ensure that AI systems are behaving ethically and responsibly.
2. *Investigating the use of AI for improving ethical decision-making:* Research could focus on developing AI systems that can assist humans in making ethical decisions, by providing information and recommendations based on ethical principles and values.
3. *Examining the ethical implications of AI-based decision-making:* As AI systems become more prevalent in decision-making processes, it will be important to understand the ethical implications of these systems. Research could focus on the potential biases and unintended consequences of AI-based decision-making, as well as ways to mitigate these issues.
4. *Developing approaches for ensuring responsible AI development and deployment:* Research could focus on developing best practices and guidelines for responsible AI development and deployment, including methods for ensuring that AI systems are transparent, accountable, and fair.

In terms of who should conduct this research, it is likely that both humans and AI could play a role. Humans could contribute their expertise and understanding of ethical principles and values, while AI could be used to automate certain aspects of the research process and provide additional insights through data analysis and machine learning.

References

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573-1576.

Chen, Y., Li, Y., Sun, Y., & Zhu, Q. (2017). Prioritizing acute myocardial infarction patients using artificial intelligence. *Scientific reports, 7*(1), 1-8.

Crawford, K., Prud'hommeaux, E., & Schultz, J. (2019). Artificial intelligence in policy-making: A tool for accountability or a new form of governance?. *Big Data & Society, 6*(1), 2053951719846187.

Crawford, K., & Schultz, J. (2017). Artificial intelligence and the double ethical bind. In *The Cambridge handbook of artificial intelligence* (pp. 559-576). Cambridge University Press.

Crawford, K., & Schultz, J. (2014). The use of artificial intelligence in employment discrimination and employment law. *Ethics and Information Technology, 16*(3), 159-172.

Dent, M. (2018). The ethics of autonomous weapons systems. In *The ethics of artificial intelligence* (pp. 95-108). Cambridge University Press.

Dignum, F., Dignum, V., & Dignum, M. (2019). *AI for a better society: Ethics, governance and accountability.* Cham, Switzerland: Springer.

Floridi, L., & Taddeo, M. (2018). *The Cambridge handbook of artificial intelligence and society.* Cambridge University Press.

Floridi, L., & Cowls, J. (2016). The ethics of artificial intelligence. In *The Cambridge handbook of artificial intelligence* (pp. 117-147). Cambridge, UK: Cambridge University Press.

Khera, O., Kocak, M., Kocak, A., & Kocak, M. (2020). Artificial intelligence in social work: A systematic review. *Social Work in Public Health, 35*(5), 469-480.

Kiritchenko, S., Zhu, X., & Cherry, C. (2018). Examining gender bias in an open-domain information extraction system. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2929-2939).

Melville, N., & Wu, X. (2019). Artificial intelligence and machine learning in supply chain management: A review and future directions. *Journal of Business Logistics, 40*(1), 11-25.

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

Topol, E. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44-56.

Addendum:
Prompts used for generating the chapter flow texts

2. Introduction to AI Ethics

Prompt: Please write an introductory chapter about what we mean by AI ethics and focus on the different meanings it can have. Provide academic references for each sentence.

3. The potential contribution of AI to AI ethics

Prompt: Describe how AI can contribute to the field of AI ethics and focus on the positive as well as negative factors. Provide academic references for each sentence.

4. Managing the risks of ethical AI biases

Prompt: Please explain how a negative ethical bias in artificial intelligence can be managed and reduced. Provide academic references for each sentence.

5. Finding the right place for AI in an ethical framework

Prompt: When should AI be employed for AI ethics and how can AI be beneficial for an ethical framework? Provide academic references for each sentence.

6. Recommendations for the employment of AI in ethical decision-making

Prompt: Please make recommendations for when AI should be used to make ethical decisions. Make individual recommendations for the fields of academia, business, health care, social work, and politics. Provide academic references for each sentence.

7. The unique contribution of AI to ethics

Prompt: What unique contribution can AI bring to ethics that would be unlikely to surface from a purely human contribution? Provide academic references for each sentence.

8.1. Summary

Prompt: Make a summary on why AI should or should not be used in the domain of AI ethics.

8.2. Future research directions

Prompt: Please make suggestions for future research directions concerning the use of AI in AI ethics and tell me if humans or AI should conduct the research.