

Alignment of Artificial Intelligence with Human Values

Parv Sharma^{1*}, Vikas Sharma¹, Priyanshu Kumawat¹

Abstract

Artificial intelligence (AI) should be directed towards humane ends. The advent of AI has created nice uncertainties about guaranteeing AI alignment with humans' worth (AI value alignment) through AI operations from style to use. For the purposes of addressing this problem, we tend to adopt the phenomenological theories of fabric values and technological mediation as a starting point. In this paper, initially, we will discuss the AI value alignment from the relevant AI studies. Second, we concisely describe what material values and technological mediation are and replicate the AI value alignment through the lenses of those theories. We tend to conclude that a set of finite human values is often outlined and tailored to the stable life tasks that AI systems will be known to accomplish. The alignment of AI values can even be fostered between designers and users through technological mediation. Upon that foundation, we tend to propose a collection of common principles to grasp AI value alignment through phenomenological theories. This paper contributes the distinctive information of phenomenological theories to the discourse on AI's alignment with human values.

Keywords: Artificial intelligence, AI automation, human values, machine learning, phenomenological theories, sustainable software development

INTRODUCTION

How Artificial Intelligence Can Help Us Better Understand Human Values

In this article, we will explore how artificial intelligence (AI) can help us better understand human values. We will discuss how AI can be used to identify and quantify values and how it can help us better understand the trade-offs we make between different values [1]. We will also look at how AI can help us make better decisions about what values to pursue.

How AI is Helping Us Better Understand Human Values

Some argue that AI is a threat to human values as it could lead to the emergence of super-intelligent machines that could make decisions that are not in line with our values [2]. However, others argue that AI can actually help us to better understand and uphold human values.

*Author for Correspondence

Parv Sharma
E-mail: 2021pietadparv034@poornima.org

¹Student, Department of Artificial Intelligence & Data Science,
Poornima Institute of Engineering and Technology, Rajasthan,
India

Received Date: April 19, 2023

Accepted Date: April 21, 2023

Published Date: May 15, 2023

Citation: Parv Sharma, Vikas Sharma, Priyanshu Kumawat.
Alignment of Artificial Intelligence with Human Values.
Journal of Law of Torts and Consumer Protection Law. 2023;
6(1): 18–20p.

For example, AI can help us better understand the complexities of human decision-making and how different values can conflict with each other. AI can also help us identify values that we may not have been aware of, and develop better ways of measuring and pursuing them.

How AI is Being Used to Help us Better Understand Our Own Values

In recent years, AI has increasingly been used to help us better understand our own values. By analyzing data related to our behavior, AI can identify patterns that reveal what we care about and

how we make decisions. This insight can be used to develop better policies and products, and to create more tailored experiences that better align with our individual values [3].

For example, AI is being used to study how people make ethical decisions. By understanding the psychological factors that influence our ethical decision-making, AI can help us make more informed choices about the right thing to do in difficult situations. AI can also help us identify our personal values, so that we can be more aware of them and make choices that are in line with them.

In addition to helping us better understand our own values, AI can also help us understand the values of others. By analyzing large datasets, AI can identify patterns in how different groups of people make decisions [4–7]. This insight can be used to develop more effective outreach and engagement strategies that are tailored to the values of specific groups.

AI is also being used to study the impact of values on social media. By understanding how values influence the content that people share and engage with on social media, AI can help us create a more positive and productive online environment.

HOW AI IS HELPING US TO BETTER UNDERSTAND THE MEANING OF VALUES

In a world where technology is increasingly becoming a part of our everyday lives, it is no surprise that AI is also playing a role in helping us better understand human values.

Through AI, we are able to gain insights into the different meanings that people attach to values such as freedom, equality, and justice. This understanding can help us create fairer systems and make better decisions that reflect the true diversity of our society.

For example, researchers at Vanderbilt University used AI to study how people from different cultures think about fairness. They found that people in individualistic cultures, like the United States, tend to value fairness as meaning that everyone should be treated the same. In collectivist cultures, like China, people tend to value fairness as meaning that everyone should be treated according to their role in society.

This research can help us create systems that are fairer to everyone, no matter where they come from. It can also help us avoid making decisions that would be seen as unfair in other cultures.

AI is also being used to help us better understand human emotions. Emotion recognition algorithms are being used to study how people react to different situations [8]. This research can help us design products and services that better meet our needs and avoid making decisions that would cause emotional distress.

THE IMPLICATIONS OF AI FOR THE FUTURE OF HUMAN VALUES

In recent years, there has been a growing interest in the potential implications of AI for the future of human values. As AI technology continues to advance, it is becoming increasingly capable of completing tasks that have traditionally required human intelligence, such as reasoning, natural language understanding, and problem solving. This raises important questions about the future role of AI in society and the impact it may have on our values.

How AI can help us better understand human values by providing us with a different perspective. In a world where technology is increasingly becoming a part of our everyday lives, it is no surprise that AI is also becoming more prevalent. With AI, we are able to get different perspectives on the world and how it works [9, 10].

One way AI can help us better understand human values is by providing us with a different perspective. For example, let us say you are trying to decide whether or not to go on a date with someone you met online.

CONCLUSION

In conclusion, artificial intelligence can help us better understand human values by providing us with new ways of looking at data and by helping us to identify new patterns. Additionally, AI can help us create new technologies that can address the problems that we face in our society.

REFERENCES

1. Stuart Russell. Aligning AI with human values. Available at: <https://arxiv.org/abs/1606.06565>.
2. Francesca Rossi, Nicholas Mattei. Towards aligning human and machine values in AI. Available at: <https://arxiv.org/abs/1805.02237>.
3. Dylan Hadfield-Menell et al. Value alignment in cooperative inverse reinforcement learning. Available at: <https://arxiv.org/abs/1606.03137>.
4. Andrew Critch, Stuart Armstrong. Value alignment problems in AI ethics. Available at: <https://arxiv.org/abs/1706.06311>.
5. Nick Bostrom. Aligning superintelligence with human values. Available at: <https://nickbostrom.com/superintelligence.pdf>.
6. Dylan Hadfield-Menell et al. Foundations for aligning AI with human values. Available at: <https://arxiv.org/abs/1711.05042>.
7. Paula Boddington. Towards a code of ethics for artificial intelligence. Available at: https://doi.org/10.1007/978-3-319-70284-1_1.
8. Daniel Weinstock. The alignment problem in ethics and political philosophy. Available at: https://link.springer.com/chapter/10.1007/978-3-030-01446-3_3.
9. Nick Bostrom, Eliezer Yudkowsky. Ethical issues in advanced artificial intelligence. Available at: <https://www.nickbostrom.com/ethics/ai.pdf>.
10. Victoria Krakovna. AI alignment: Why it's hard, and where to start. Available at: <https://www.lesswrong.com/posts/sEucKjZLbuxhQ6XdR/ai-alignment-why-it-s-hard-and-where-to-start>.