Conscious AI Should Be Managed Similarly to Humans

Arushi Saurabh

Abstract

AI systems are becoming increasingly prominent and ubiquitous in our daily lives. For example, one can find AI systems in social media and in large language models (ex. ChatGPT) used to predict user behavior and text input. While these AI systems can be useful, they still have problems aligning with our human values. This paper will conduct a systematic review of ethical AI design methods, and discuss instances where these design methods assisted AI in aligning with human values. Then we will discuss how AI developing human consciousness may change the implementation of these design methods. This paper provides a potential way to think about how to manage AI in the case it develops human consciousness.

Introduction

The rise of AI has been accompanied by debates about the ethical implications of the advancements in AI. There are already ethical debates about AI applications like ChatGPT having the ability to write out essays and how this may impact human creative processes (Dwivedi et al., 2023). Although the concept of how to keep AI in check in its current state has been studied thoroughly, research on how AI developing consciousness may affect AI checks is limited. The question we aim to answer here is how can we ensure that AI systems are aligned with human values and ethical principles? This paper argues that AI with human consciousness should be regulated similarly to how humans are. It first talks about instances where AI does not align with human values, then evaluates current methods used to keep AI aligned with human values, and finally addresses how AI consciousness will change things in the future.

Current instances of misalignment between AI and human values

Healthcare

AI used in healthcare has been shown to have dangerous racial and gender biases that can lead to misdiagnosis (Parikh et al., 2019). These biases are caused by AI being trained on limited data sets that don't account for much diversity (Celi et al., 2022). One example of this was

when an AI was tasked at interpreting chest x-rays and displayed gender disparity. Mainly men's x-rays were uploaded onto the AI so the accuracy it had when analyzing women's x-rays was pretty low. Another AI model displayed disparities in its ability to diagnose skin cancer depending on the skin color of the subject. It was able to diagnose skin cancer more accurately for light-skinned subjects since it was mainly trained with samples from people with lighter skin. This means, when using the model on darker skinned subjects, it may either completely misdiagnose them or it won't be able to detect the skin cancer until it's too advanced (Fulmer, 2022). This violates the human values of fairness as well as accessibility to healthcare since the data used to train the AI lacks diversity therefore resulting in unequal quality of care for different racial and gender groups.

Social Media

Social media algorithms generate a feed for social media users based on what they have viewed and interacted with in the past. This can help users have a more enjoyable experience with social media that is personally catered towards them, however, it can also have some adverse effects. For example, social media has increased political participation among users but it has also caused political polarization. Algorithms will direct users to posts congruent with their beliefs which can cause hyperpartisanship. These algorithms could also spread misinformation which can lead to even more polarization (Tucker et al., 2018). Politics in social media is an example of an echo chamber. Echo chambers limit the exposure of social media users to diverse perspectives therefore causing certain opinions of theirs to be reinforced without a full knowledge on the issue (Cinelli et al., 2021). These echo chambers infringe upon fundamental human values such as freedom of thought, access to diverse information, and, in some cases, well-being, particularly when individuals are exposed to harmful content instead of professional guidance. One example of this is regarding mental health. People with similar mental health disorders may be fed content by people with the same mental health disorder as them when they should be seeing content from mental health professionals.

Large Language Models

Large language models are AI that have been trained on large amounts of text data which enables them to generate human-like responses. They have gained traction recently with the public release of ChatGPT which is now frequently used, especially in the field of education, by teachers and students alike. While ChatGPT can greatly assist teachers by coming up with lesson plans and otherwise improving their students' quality of education, it can also inhibit the creative processes of students by writing text responses for them. As students continue to

use ChatGPT for tasks meant to challenge them while they're still in the process of developing their creative thinking skills, the essential human value of creativity will face a steep decline. Many have spoken out against ChatGPT for negatively impacting students' education by taking away from their learning of essential writing skills that are considered necessary for their future. There are also fears of students' critical thinking and problem solving skills being negatively impacted by ChatGPT as well due to how simple it is to find information that would normally require some work to find (Kasneci et al., 2023).

Evaluation of Design Methods

Value Sensitive Design

Value sensitive design (VSD) accounts for human values in the design process. It places an emphasis on ethics and morality when designing AI. VSD also acknowledges the impact that technology has on human lives and ensures that impact is positive by considering ethical implications early on in the design process. VSD uses a tripartite methodology that uses conceptual, empirical, and technical investigations. To grasp a human perspective while designing, VSD involves stakeholders that are strongly affected by the technology they are designing on (Friedman & Hendry, 2019; Friedman et al., 2002). Value sensitive design was used during the COVID-19 pandemic to help minimize the spread of the virus. The Robert Koch Institute (RKI), the German research facility for disease control and prevention, created an app called Corona Datanspende (Corona Data Donation) for users to voluntarily share their health data so they could track the spread of COVID. At the beginning of the design process, VSD was used to identify what values the design was meant to promote as well as values it must respect for the wellbeing of its users. The values that the design was meant to promote aligned with the UN's third sustainable development goal of "good health and wellbeing." The values the design was meant to respect were human autonomy, prevention of harm, fairness, and explicability. After values were determined, the designers used visualization to determine technical design requirements. Then, prototyping was used to determine if the design actually aligned with these values and also assessed the behavioral impacts of it. This case in particular revolves around a product that has the potential to add to the understanding of the COVID-19 virus, which would be important for the health of the German citizens, however, its use of personal data could cause some ethical issues. This product's potential impact on citizens of Germany makes it a very important case study into value sensitive design. Keeping users' and designer's values in mind during the design process ensures the product will achieve its goal and be as user friendly as possible while also respecting ethical standards (Umbrello & Van De Poel, 2021).

Participatory Design

Participatory design enables diverse stakeholders to play an active role in the future of AI. This design method is incredibly useful for keeping humans' best interests at heart when designing AI so that it can properly align with human values. Using diverse perspectives when designing AI can prevent AI from unknowingly harming marginalized groups and can minimize bias. Currently, the exact extent to how much humans are able to participate in the design isn't clearly defined so establishing guidelines for that may be useful (Zytko et al., 2022; Birhane et al., 2022; Delgado et al., 2021). In the field of law, diversity of perspectives is necessary for everyone to be adequately represented. Lawyers have a great responsibility as their work can have profound impacts on the livelihood of many people. To help benefit lawyers and their clients, the National Institute of Standards and Technologies used participatory design to connect different perspectives and create the Text REtrieval Conference's Legal Track. The Legal Track's purpose was to assist collaboration between attorneys of different sides, and was specifically designed for cases where opposing attorneys needed to share evidence. During the design of the Legal Track, real lawyers were used as stakeholders so that the product would be specifically catered towards their needs. The designers of the Legal Track introduced a special role called "Topic Authority" (TA) to bring stakeholders into the loop during the design process. The designers had been having a problem with filtering relevant evidence using the algorithm so incorporating stakeholders to determine the relevancy of the information was very useful. The TA position also assisted in finding human errors during document review. Overall, the interactive environment fostered through participatory design connected lawyers with computer scientists to design a beneficial AI system that could work in high-stakes scenarios (Delgado et al., 2022).

Algorithm Audits

Algorithm auditing is a process where an auditor looks at the system and finds issues with it then makes recommendations to the designers on how to repair those issues. However, designers of algorithm technologies can say their product has been audited when it actually hasn't, which can be a problem if the system has clear issues. Therefore, there needs to be more auditing regulations put in place because when done appropriately, algorithmic auditing is a necessary part of the design process and can vastly improve the equity of systems (Costanza-Chock et al., 2022). A startup called pymetrics used algorithmic auditing for their product that used machine learning to assess which applicants would be best suited for a job and move them up to the interview stage. Pymetrics assessed which applicants were best for the job by comparing the applicant's gameplay of pymetrics'

suite of games to the gameplay of a high performing, incumbent employee. The audit of pymetrics focused mainly on assessing its algorithm's correctness, discrimination, and de-biasing circumvention. The audit found that there were no issues with these three aspects of the algorithm. This was an example of a cooperative audit where the company and the auditors set specific rules for what exactly the auditors would assess. For example, the auditors didn't assess whether pymetrics actually adequately found the best candidates for the job, they just made sure the algorithm worked as it was supposed to, wasn't biased, and was safe from people who may try to cause bias. Although there were ultimately no issues found with the algorithm, the audit still ensured the fairness of the pymetrics system. The job hiring process has the potential for harmful biases which makes the auditing system necessary (Wilson et al., 2021).

Combination of Methods

Value sensitive design and participatory design both must be implemented during the design process, whereas, algorithmic auditing is used after the design has been finished. All three of these methods could be used in conjunction with each other to design a product. First, a group of stakeholders would need to be found. They are necessary for both participatory and value sensitive design. Once those stakeholders are found, their values as well as the designers' values would need to be assessed. These values would drive the design of the product. After that, a prototype would be created with both the stakeholders and the designers playing a part in its creation. Once there is a solid prototype, algorithmic auditing would be done to ensure the product works as it should and has no signs of bias. Using all three of these methods in conjunction could create an effective design that was created ethically and with the users' opinions in mind.

How AI consciousness can change things

Although these checks may be sufficient for the current state AI is in, they may be unable to pass the test of time if AI develops consciousness. For an AI to have consciousness, it would have to be self aware. It must also be sentient and have subjective qualitative experiences (Hildt, 2019). In the case of conscious AI, we would have less control over the actions of AI than we do currently with non-conscious AI because it would be capable of thinking independently for itself. Therefore, we must figure out new systems to effectively manage AI. What will AI be capable of if it ever develops consciousness, and how should we enforce the regulations on AI given any new capabilities? We may have to develop a punitive system, similar to what we have to manage humans, with laws and punishments. AI regulations are currently centered around data privacy. These regulations are mainly placed on companies or other

manufacturers of AI products to ensure that AI is used ethically (Hoffmann-Riem, 2020). However, if AI gains consciousness, those regulations will be placed on the AI itself because it will have the ability to control its own actions. This is similar to how humans are managed because they are obligated to hold themselves accountable for their actions. Therefore, controlling AI with human consciousness could be handled similarly to how humans are managed. With AI being conscious, humans could work hand in hand with AI throughout the design process for a product or service. In value sensitive design, AI could work as an efficient designer who could help assess the values and needs of stakeholders by analyzing data and then interpreting it similarly to humans, which non-conscious AI is incapable of. It may be able to assess the company's implicit values better than an employee because of its ability to analyze large amounts of company data. AI could also predict users' values using user data. This could eliminate the need for direct involvement from human stakeholders but initially they may still be required as a check for the AI. Since AI may not be completely reliable in the beginning, humans would still be necessary to verify the AI's work. The role of humans in the design process will shift as AI gains consciousness from being directly involved in the design process to becoming more like auditors for the AI. This hierarchy where humans are checking over the AI's work is similar to a typical workplace therefore the AI is being managed similarly to the way humans are currently managed at work. For participatory design, however, human stakeholders would still be necessary but AI could still probably stand in to imitate a human's experience with the product. Some believe that we can already make this happen using ChatGPT to stand in for human stakeholders. However, this is not possible because ChatGPT does not yet have human consciousness so it's incapable of thinking for itself and therefore will not be able to adequately imitate a user's experience. Conscious AI would be far better suited to stand in for human stakeholders since it would be able to better behave like a human since it would have a similar thinking process to humans. However, human's perspectives are still needed to check AI's potential mistakes. Humans would also still be necessary for algorithmic auditing because AI auditing other AI could further perpetuate certain biases. One current issue with AI is its inability to identify bias in its data. Currently, AI used in the realm of healthcare is greatly impacted by limited or biased data which can cause adverse results for users. But if AI were to develop consciousness it may have the potential to determine whether its data is biased since it would be able to think subjectively which could revolutionize healthcare. But since it's only trained on that data, it may not be able to recognize the flaws in its thinking. Humans are very similar in this way because they grow up in environments with very specific ways of thinking and don't recognize their thinking is biased until they go to school or leave that

environment. So to dismantle biases in AI, there may need to be an education system formed specifically for AI. Overall, all the checks that would be considered sufficient to keep AI aligned with human values currently will need to be altered if AI ever develops consciousness. As well as potentially having the ability to identify bias, AI's algorithms will become less black and white if it develops consciousness. Currently, if a social media user seems to demonstrate an agreement with a certain political party that user's algorithm will flood their feed with posts that are biased towards that side. This has caused hyperpartisanship and polarization among social media users (Barnidge & Peacock, 2019). However, if AI is able to identify biased posts and is also more aware of complexities in political opinion due to its ability to think subjectively, social media algorithms will become less biased therefore potentially bringing down the rates of political conflict on social media. The way we manage AI is subject to a lot of change if AI ever develops human consciousness. As AI gains human traits such as sentience and the ability to think subjectively it makes sense why it should also be managed similarly to humans. Developing a punitive system similar to the one for humans could be useful for keeping AI in check as well as establishing an education system and hierarchy in the workforce which are also integral to managing humans. If we can learn to effectively manage AI it can play an even larger part in design methods such as value sensitive design, participatory design, and algorithmic auditing.

Conclusion

As AI becomes a bigger part of our lives and gets closer to achieving human consciousness, we must think of how we can regulate it in the future. Currently, instances of AI not aligning with human values in healthcare and social media have to do with bias. Some methods used in the design process of AI systems can be used to mitigate these biases. These methods are value sensitive design, participatory design, and algorithmic auditing. These methods can be combined to account for human values throughout the design process and prevent AI from misaligning from standard ethical principles. However, if AI were to develop consciousness, changes would have to be made to how we regulate it. A system, similar to how humans are currently managed, should be created to manage AI. This paper's main limitation is that it is highly hypothetical as the issue of AI achieving human consciousness is very uncertain and there isn't any concrete data concerning it.

References

Barnidge, M., & Peacock, C. (2019). A Third Wave of Selective Exposure Research? The Challenges Posed by Hyperpartisan News on Social Media. *Media and Communication*, *7*(3), 4–7.

https://doi.org/10.17645/mac.v7i3.2257

Birhane, A., Isaac, W. S., Prabhakaran, V., Díaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. arXiv (Cornell University). https://doi.org/10.1145/3551624.3555290

Celi, L. A., Cellini, J., Charpignon, M., Dee, E. C., Dernoncourt, F., Eber, R., Mitchell, W., Moukheiber, L., Schirmer, J., Situ, J., Paguio, J. A., Park, J., Wawira, J., & Yao, J. S. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. PLOS Digital Health, 1(3), e0000022. https://doi.org/10.1371/journal.pdig.0000022

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. Proceedings of the National Academy of Sciences of the United States of America, 118(9). https://doi.org/10.1073/pnas.2023301118

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. 2022 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3531146.3533213

Delgado, F., Barocas, S., & Levy, K. (2022). An Uncommon Task: Participatory Design in Legal AI. Proceedings of the ACM on Human-computer Interaction, 6(CSCW1), 1–23. https://doi.org/10.1145/3512898

Delgado, F., Yang, S. C., Madaio, M. P., & Yang, Q. (2021). Stakeholder Participation in AI: Beyond "Add Diverse Stakeholders and Stir." arXiv (Cornell University). https://doi.org/10.48550/arxiv.2111.01122

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Friedman, B., & Hendry, D. G. (2019). Value Sensitive Design:

Shaping Technology with Moral Imagination. MIT Press.

Friedman, B., Kahn, P. H., & Borning, A. (2002). Value Sensitive Design: Theory and Methods. UW CSE.

Fulmer, J. (2022). Addressing AI and Implicit Bias in Healthcare. TechnologyAdvice. https://technologyadvice.com/blog/healthcare/ai-bias-in-healthcare/

Hildt, E. (2019). Artificial Intelligence: Does Consciousness Matter? Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.01535

Hoffmann-Riem, W. (2020). Artificial Intelligence as a Challenge for Law and Regulation. Springer eBooks, 1–29. https://doi.org/10.1007/978-3-030-32361-5_1

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. JAMA, 322(24), 2377. https://doi.org/10.1001/jama.2019.18058

Tucker, J. A., Guess, A. M., Barberá, P., Vaccari, C., Siegel, A. A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. Social Science Research Network. https://doi.org/10.2139/ssrn.3144139

Umbrello, S., & Van De Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. AI And Ethics, 1(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3

Wilson C., Ghosh A., Jiang S., Mislove A., Baker L., Szary J., Trindel K., and Polli F.. (2021). Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In ACM Conference on Fairness, Accountability, and Transparency https://doi.org/10.1145/3442188.3445928

Zytko, D., Wisniewski, P., Guha, S., Baumer, E. P. S., & Lee, M. G.

(2022). Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. CHI Conference on Human Factors in Computing Systems Extended Abstracts. https://doi.org/10.1145/3491101.3516506