Yagna Patel

Mr. Barrett

English - H225

May 06 2023

What ethical guidelines and principles should be used in the development and use of Natural Language Processing (NLP) powered AI and algorithms, and how can we ensure that these AI implementations reflect diversity of perspectives and values?

Table of Contents

Introduction

The world is rapidly changing with new emerging technologies. These growing technologies are affecting our day to day lives at a higher scale than ever before. One of the most recent, and perhaps more significant, technologies is A.I., or Artificial Intelligence. Just as the name suggests, A.I. refers to the development of computer systems that can perform tasks that typically require human intelligence, such as learning, reasoning, and problem-solving. A specific type of A.I. that is set to revolutionize society is Natural Language Processing (NLP) powered A.I. This type of A.I. has the potential to transform the way we communicate, process information, and make decisions. NLP powered AI and algorithms allow machines to comprehend, analyze, and produce human-like language. This can be applied to a wide range of applications, like chatbots, virtual assistants, language translation, and even to sentiment analysis. Natural Language Processing (NLP) tools, such as ChatGPT, Bard, and Whisper AI, have gained widespread popularity in recent years due to their ability to increase productivity efficiency and streamline tasks for businesses and individuals. However, the use of NLP-powered A.I. also raises important questions about what ethical guidelines and principles should be used to guide the development and use of NLP-powered AI and algorithms, and how we can ensure that these AI implementations reflect a diverse perspective.

Methodology

To fully examine the ethical implications of NLP-powered AI and algorithms, this study takes a mixed-methods approach, which includes a literature review, empirical testing, and an expert interview. The research begins with a thorough literature review that involves searching for academic articles, books, and reputable reports. The literature review focuses on exploring

key ethical principles such as diversity, data privacy and security, open source vs closed source, and non-discrimination in the context of NLP development. In the second phase of the research, two widely recognized NLP models, ChatGPT and BARD, undergo empirical testing. The testing is conducted across various situations, including sensitive topics, to assess their responses and determine any potential limitations or biases. The testing process examines their outputs, with an emphasis on categories like gender, race, politics, and religion. The research seeks to reveal any failures or biases in the algorithms of these NLP models. In the third phase of the research, an expert in the field is interviewed to gain valuable insights into NLP development and research. Professor Gladbach from the University of Missouri - Kansas City, who possesses extensive knowledge in computer science and algorithms will be interviewed. The study aims to gain a deeper understanding of challenges faced in the field by incorporating expert opinions and personal experiences. Overall, the combination of literature review, empirical testing, and an expert interview will help to analyze the ethical implications of NLP-powered AI and algorithms.

Ethical Guidelines

Natural language processing technology offers numerous benefits, but it also poses potential harms that must be addressed. One major concern for NLP algorithms is that they may reinforce existing biases leading to discrimination and social inequalities (Kocaballi et al., 2020). Additionally, the use of NLP technology can raise privacy concerns because it involves the collection and processing of large amounts of personal data (Chen et al., 2021). Accountability and transparency are another significant hurdle to overcome when it comes to the development of and utilization of NLP/AI. Without clear accountability mechanisms and transparent reporting of outcomes, it can be difficult to identify and address errors or biases in the technology (Bender

et al., 2021). This not only leads to distrust from users and stakeholders but also lessens the effectiveness of the technology. In a recent test conducted on OpenAI's ChatGPT model, it was found that the AI keeps track of a wide range of information,

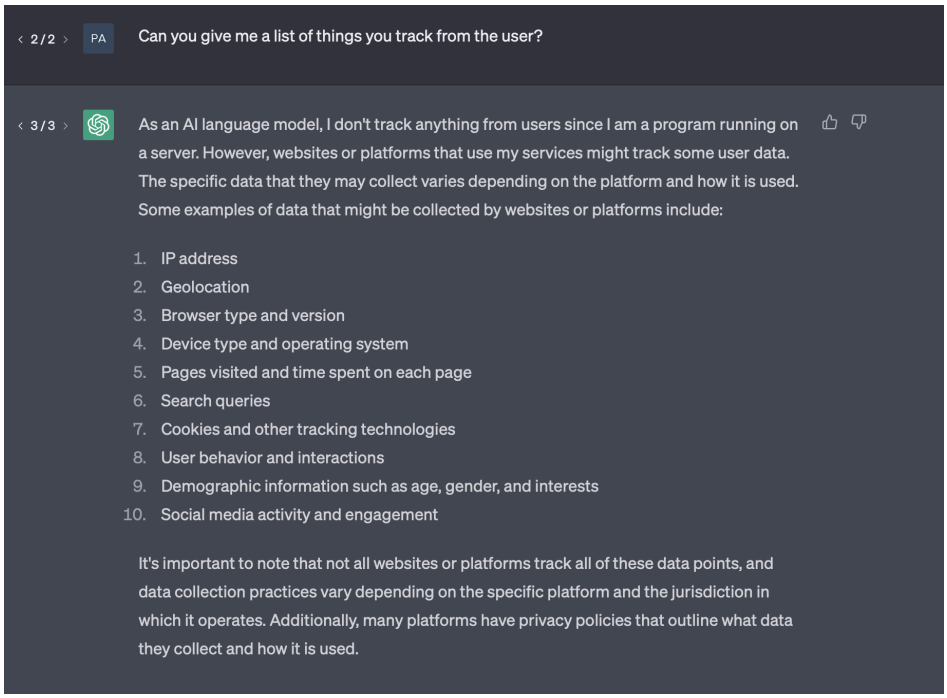

Figure 1.

Figure 1 illustrates the need for proper measures to safeguard data as a leak or wrong use of data can each be catastrophic not only for the company but also for the public in general. Additionally, neglecting guidelines is an issue that needs to be addressed as it is quite relevant in the tech industry. Every major tech company collects data and sells it. And many times it is utilized unethically. Consequently, the use of NLP technology brings up ethical concerns, particularly in areas like privacy, bias, and fairness. The further development and use of NLP technology can lead to unintended consequences, such as social inequalities or undermining trust in institutions (Grafe et al., 2021). Therefore it is important to develop and follow clear ethical

guidelines that prioritize the needs and rights of users and society as a whole instead of corporate interest.

Key Ethical Principles for NLP Development

NLP algorithms are developed from vast amounts of data, making it crucial for developers to follow principles such as diversity, data privacy and security, and non-discrimination. It is important to ensure that the representation of different communities and cultures is reflected in the data used for training NLP models. According to Liu "biases in NLP models can be amplified and perpetuated if the data used to train them is not diverse enough to capture the nuances of different groups" (Liu et al. 204). Therefore, including diverse perspectives in the development and NLP systems will help to ensure higher likelihood of inclusiveness and fairness in the implementation of NLP models. This can be achieved by using data from a wide range of sources and involving stakeholders from different backgrounds in the design and testing of NLP systems. Data privacy and security are just as important. The use of large amounts of data to train NLP models raises concerns about data privacy and security. To keep a good balance there are only two options: Open source or Closed source. Open-source data can be vulnerable to attacks, but with that downside comes the benefit of community support and transparency. A great example is Wikipedia, which allows for a large community support, but at the cost of misleading information. Unlike open source the use of closed-source data can limit transparency and accountability but has the added benefit of security, technical support, and financial support (GeeksforGeeks, 2021). Professor Gladbach at the University of Missouri - Kansas City expressed her concern with a personal story when asked about closed source vs open source, "I think about the Tesla my son has….If the software was open source – I am concerned for the ability for anyone to 'hack' into the car's control devices." (Interview). To

address these concerns, NLP developers must secure the users using data that is collected and store it in a secure manner. This includes obtaining informed consent from participants and implementing strong data protection measures. Non-discrimination is a critical ethical principle in NLP development. It is important to ensure that NLP models do not perpetuate or amplify biases based on race, gender, ethnicity, or other protected characteristics. To prevent discrimination, NLP developers must use diverse data and take steps to identify and mitigate biases in their models. This may involve conducting bias audits, testing for fairness, and implementing algorithms that promote inclusivity and equity. Ultimately, the goal is to develop NLP systems that reflect the diversity and complexity of human communication while ensuring that they are ethical and equitable for all users.

Ensuring Diverse Perspectives and Values in Ethical Guidelines

Diverse perspectives and values are crucial in developing equitable and inclusive NLP systems. One way to ensure diversity is by involving developers, researchers and even stakeholders from diverse backgrounds in the development. This can include experts in fields such as linguistics, sociology, and psychology, as well as community representatives and marginalized groups because “participatory approaches to ethical decision-making can ensure that the needs and perspectives of diverse communities are considered” (Caliskan et al. 382). Including diverse perspectives can help ethical guidelines better reflect the values and needs of all users and prevent bias and discrimination. As a colleague of Professor Gladbach noted, “ChatGPT acts like it knows all, but sometimes provides very wrong responses and makes up references” (interview). This emphasizes the importance of including experts and community representatives from different fields and backgrounds in the development of ethical guidelines to ensure that all voices are heard and considered. Through my testing of ChatGPT and BARD,

another NLP AI trained by Google, I found that both of their algorithms avoid answering questions involving racism, religion, sexuality, and politics. However, what was interesting to note was that neither of the NLP models were able to 'compare' subjective aspects of those controversial topics. Moreover, when I asked it to tell me a joke about Secularism, both were able to do so, but when I told both to tell me a joke about 'wahhabi state' it responded with "I cannot generate offensive content". This emphasizes the fact that both are trained to avoid generating content that could be offensive has failed, meaning there needs to be more alertness in the development through ethical guidelines. Ultimately, incorporating diverse perspectives can lead to the creation of more inclusive and equitable NLP systems but even then there will be small places where AI can be exploited in the system thus we have to make sure to keep it safe.

Conclusion:

In conclusion, the development and utilization of NLP-based AI and algorithms carry significant ethical considerations that require attention to ensure they are ethical, fair, and inclusive for all users. Key ethical principles, including diversity, data privacy and security, and non-discrimination, must be integrated into the development and use of NLP systems to prevent bias and discrimination. Additionally, incorporating diverse perspectives and values is critical to reflecting the needs and values of all users in ethical guidelines. As NLP technology advances, developers must prioritize the needs and rights of users and society as a whole over corporate interests. By doing so, we can maximize the potential benefits of NLP-powered AI and algorithms while minimizing their potential harms. Ultimately, all stakeholders, including developers, regulators, and users, must collaborate to develop and implement ethical standards.

Further Research:

The world of AI is vast, with numerous subdivisions and many different applications. Training AI requires a significant amount of power, utilizing multiple Graphical Processing Units and Computing Processing Units, which consume large amounts of energy resources. Therefore, taking a closer look at the effects of AI on climate change would be an interesting research topic. By analyzing the energy consumption of AI technologies, researchers can identify ways to reduce the carbon footprint of these systems and ensure that their development and use align with sustainability goals.

Another area of interest is researching the military applications of AI and its potential for weaponization. The use of AI in military settings raises important ethical and political questions, especially when it comes to the use of autonomous weapons and the possibility of human rights abuses. By examining the impact of AI on military operations and international relations, researchers can identify potential risks and benefits of these technologies. They can also propose ethical guidelines and regulations for their development and use in military contexts. This research can help ensure that AI is developed and used in ways that align with international laws and ethical principles. It can also promote peace and security in the world.

Resource Inequality is another issue as AI technologies progresses. It increases the need for computing power, data storage, and energy resources. This can create a gap between those who have access to these resources and those who do not, making the already existing inequalities even worse. Smaller companies, developing countries, and economically disadvantaged individuals may find it difficult to access or utilize the full potential of AI because of the high costs and resource requirements. To tackle this issue, we need to find ways to make AI technology more accessible and affordable.

Works Cited

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. arXiv preprint arXiv:2102.06171.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2020). Ensuring AI benefits diverse stakeholders requires avoiding moral outsourcing. Communications of the ACM, 63(6), 380-389.

Chen, R., Ge, Y., & Han, J. (2021). Natural language processing and big data. Foundations and Trends in Information Retrieval, 15(2-3), 95-276.

GeeksforGeeks. "Difference Between Open Source Software and Closed Source Software." GeeksforGeeks, 15 Sept. 2021, https://www.geeksforgeeks.org/difference-between-open-source-software-and-closed-source-software/.

Grafe, S., Wirth, J., & Gaisser, S. (2021). Ethics of Artificial Intelligence: The Intersection of Law, Technology, and Society. Oxford University Press.

Kocaballi, A. B., Laranjo, L., Coiera, E., & Eysenbach, G. (2020). The potential of social media for health-related research and surveillance. American Journal of Public Health, 110(S3), S249-S253.

Liu, J., Sap, M., Gabriel, S., Smith, N. A., & Choi, Y. (2021). On the Danger of Bias Amplification in Natural Language Processing. arXiv preprint arXiv:2103.10385.

United States. Executive Office of the President. "The National AI Strategy: 2022 Annual Report to Congress." 5 December 2022, https://www.whitehouse.gov/wp-content/uploads/2022/12/TTC-EC-CEA-AI-Report-12052022-1.pdf.