Ethics and Social Impact of AI Tutors

Qumeng Sun
University of Goettingen
Goettingen, Germany
sun.qumeng@stud.uni-goettingen.de

***Abstract*—The application of Artificial Intelligence (AI) as teaching assistants in education has shown significant potential for personalizing instruction and improving educational effectiveness. However, the ethical and social implications of this technology necessitate careful examination. This study reviews AI in Education (AIED) literature, revealing potential ethical challenges AI teachers may face. It further discusses the effectiveness of AI teachers and the need for proactive interventions, responsiveness, community values, and a practical regulatory framework to ensure ethical use. The study concludes that while AI teachers can significantly benefit education, proactive measures are necessary to address potential ethical and social issues, ensuring that technology is a beneficial tool rather than a source of problems.**

Keywords—Artificial Intelligence in Education (AIED), Ethical Challenges

I. Introduction

In recent years, the application of Artificial Intelligence (AI) in education has made significant progress [2,3,4], and one important application is AI as a teaching assistant. Traditional human teaching assistants have limitations in providing personalized instruction to meet the needs of each student. In contrast, AI teachers, as an emerging teaching aid, can personalize instruction and guidance based on a student's learning progress and level of understanding through advanced technologies and algorithms. This personalized teaching method has been proven to improve students' academic performance and educational effectiveness significantly [17,18,19].

By reading papers on Artificial Intelligence in Education (AIED) and incorporating existing regulations, we will reveal the ethical issues and challenges that AI teachers may face. This thesis will review relevant literature and research to establish a theoretical foundation for applying AI teachers as teaching assistants. Next, we will introduce the research methodology and data collection process to ensure the reliability and validity of the study. Then, we will analyze the effectiveness and impact of AI teachers in personalizing instruction and discuss possible ethical issues. Finally, we will summarize the findings and present relevant conclusions and recommendations to promote the sustainable development and application of AI teachers in education.

By reading papers on Artificial Intelligence in Education (AIED) and incorporating existing regulations, we will reveal the ethical issues and challenges that AI teachers may face. This thesis will review relevant literature and research to establish a theoretical foundation for applying AI teachers as teaching assistants. Next, we will introduce the research methodology and data collection process to ensure the reliability and validity of the study. Then, we will analyze the effectiveness and impact of AI teachers in personalizing instruction and discuss possible ethical issues. Finally, we will summarize the findings and present relevant conclusions and recommendations to promote the sustainable development and application of AI teachers in education.

II. AI in Education

A. AI Tutor

Teaching assistants(TAs) are a very effective way of learning. Students who learn through one-on-one tutoring are likelier to score higher than in traditional education [1]. AI tutors are particularly well suited for this role.AI tutors can provide personalized instruction, adapting to each student's learning progress and level of understanding, which is difficult for traditional human tutors to do [2,3].

B. Applications

A natural first way to bring AI teachers into the classroom is to use them as teaching assistants. OpenAI has partnered with Khan Academy to support its efforts to create fine-tuned models based on GPT-4 to help students learn [2].

This chatbot called Khanmigo is very powerful. Not only can it answer questions posed by students, but it can also give them specific guidance based on their thoughts and guide them closer to the correct answer when they make mistakes instead of pointing them out rigidly. According to Khan Academy holders, the bot's critical ability is to "ask each student personalized questions to promote deeper learning."

EdTech companies, including Duolingo and Quizlet, have integrated OpenAI's chatbot into their applications [21]. Universities belonging to the Russell Group, such as Oxford, Cambridge, and some other top universities, have begun to allow and teach students how to use GPT in their studies [23]. The use of AI in education has been encouraged by mainstream educational institutions [3]. The GPT-4 has been used as a suitable assessor in several studies [20], and their use in various writing tasks can increase productivity [38]. It can generate questions of varying difficulty, multiple-choice questions, and even clinical exam scenarios [19]. Such applications contrast particularly nicely with previous concerns about cheating on the GPT-4 [3, 22]. However, the use of AI robots in education does require caution. In addition to academic integrity, there are many potential ethical and social issues of concern.

III. Ethics in the AI Model

In this section, the paper looks at three potential ethical issues AI teachers may exhibit and suggests possible solutions.

A. Truthfulness

Artificial Intelligence (AI) has the potential to revolutionize numerous sectors, but it also poses significant ethical challenges, one of which is authenticity. AI models can generate highly realistic synthetic data, such as Deep-Fake, which can be used to deceive individuals and manipulate public opinion [24]. The ability of AI to create such convincing false information raises serious questions about trust and authenticity in the digital age. It is crucial to develop robust verification mechanisms to ensure AI-generated content's authenticity and prevent misuse [25].

Large language models can generate responses quickly, but these responses are only sometimes truthful. This is because these models generate text by predicting the next word; they cannot guarantee the truthfulness of the information generated, only the fluency of the language [29-34, 38, 39]. They could improve in some education-related tasks, such as math. For addition up to two digits, they can have a high correctness rate, but for more complex tasks, the correctness rate drops sharply to nearly 10% [35]. Misleading information is unacceptable for AI teachers and is one of the most critical issues. Currently, researchers are trying to improve or limit the content generated by the model for a specific topic through various methods, such as using particular search strategies and reinforcement learning to help AI overcome this challenge in math [7,8].

B. Privacy[4]

Another ethical concern in AI is the protection of privacy. AI systems often rely on vast amounts of data, some of which may be personal or sensitive. Using such data can lead to improper management of privacy violations [26]. For instance, AI algorithms used in healthcare can reveal sensitive patient information if not adequately anonymized[27]. Therefore, it is essential to implement robust data protection measures in AI systems, including data anonymization and secure data storage and transmission methods.

In education, student privacy protection has always been an important topic. For research on extensive language modeling, we should ensure that students' data are secure and that no sensitive information flows into the model and the cloud. Several studies have pointed out this concern for data security [40, 41]. In the application of AI teachers, handling students' data and how to protect students' privacy have become imperative [36]. The New South Wales Department of Education, in a survey published in 2020 [37], listed eight points of principles that AI should follow. Similar regulations limit the behavior of researchers during data collection and experimentation, ensuring that they collect data legally and ethically. In addition, AI teachers should be private to the students or the school, restricting others from accessing and processing the data. [41] argues that using a perfect database for training is the most straightforward solution. Other methods against data leakage are also effective.

C. Harmful information

AI models can also inadvertently generate or propagate harmful information. For instance, AI algorithms used in social media platforms can amplify harmful content, such as hate speech or misinformation, by promoting content that generates high user engagement, regardless of its truthfulness or potential harm [28]. This can have profound societal implications, including the spread of fake news, polarization, and radicalization. Therefore, developing AI systems that can effectively identify and mitigate the spread of harmful information is crucial.

AI models may output harmful information [15], which is especially important to be aware of when applying AI teachers. We must ensure that AI does not output any toxic information, as this may lead to psychological harm or reinforce harmful stereotypes in students. Currently, most social media resist damaging information [44] through different ways, including lousy word filters, AI detectors, etc [45, 46]. We can better train harmful information detectors by manual annotation or semi-supervised datasets. A common way is adversarial attacks [15, 42]. Perspective API is the most widely used API for detecting toxic information [43]. FM Detector [6] is recommended to cope with such harmful information in a review study on the underlying model. The idea is also to use AI to counter false information generated by AI. Training similar classifiers relies on relevant databases, and ethical use of these data for abusive content such as hate speech and harassment is also challenging [14].

In conclusion, despite its potential for significant benefits, AI also raises serious ethical issues that must be addressed. Ensuring the authenticity of AI-generated content, protecting privacy, and preventing the dissemination of harmful information are vital ethical challenges that need to be discussed during the development and deployment of AI systems. In addition to the models' ethnicization, we need to ensure that this AI is used correctly in society.

IV. Ethics outside the AI model

This section will discuss the impact of AI models on society.

A. Fairness and Justice

1) Educational Equity

Equity and justice are important issues we must consider when considering AI teachers. In education, uneven distribution of resources is a persistent problem [47]. Educational resources often tend to be concentrated in large cities, wealthy and majority-ethnic families. For example, sixth-grade students in the most affluent districts in the United States outperform students in the poorest districts by four grades [49]. Educational inequities are even more pronounced globally. In the last decade, some top universities have used free and open courses to lower the walls of knowledge and create a competitive advantage for themselves [48]. But there are some problems, [50] arguing that this has exacerbated other inequalities, such as course selection being more likely to focus on courses with national faculty. Building on this, this thesis argues that the emergence of AI teachers may change this situation further. As AI is static, once it is trained, it can be used at no extra cost by simply paying for electricity.AI could replace some of the work of teachers and help provide more affordable and high-quality basic education, especially for those at the bottom of the social ladder, who could benefit to a greater extent from the improved quality of education brought about by AI teachers [51].

2) AI & Capitalism?

We also need to be wary of AI becoming an expensive commercial product. People who use better AI may gain additional advantages in work and study [52]. Historical discrimination may be exacerbated in society due to the unequal distribution of resources [5, 6]. This may continue to happen in the age of AI. Currently, giants such as OpenAI, Microsoft, Google, and Amazon hold the majority of computational resources, and it is largely the responsibility of these large corporate entities to push the frontiers of the underlying models [55]. This could lead to a significant concentration of decision-making and power, reducing income and opportunities for those who have ownership. This concentration of power may lead to an equilibrium where fewer people have social and economic mobility and opportunity [53, 55].

3) Misguided Values

Thought control or misdirection of values may occur among AI teachers. The relationship between education and brainwashing is delicate; both involve transferring knowledge and ideas. As with any AI system, AI teachers may exacerbate existing inequalities by producing unfair outcomes, entrenching systems of power, and disproportionately distributing technology's negative consequences to those already marginalized [5, 54]. Such issues help amplify viewpoint homogenization and monocultures, which usually emerge with overrepresentation bias. For example, Zhou's paper in 2021 states that they found that words that occur less frequently in BERT, one of the most cited current language models, lead to a significant performance gap. From this, they argue that this imbalance in the data leads to advantages for a specific culture.

4) Solutions

To address the issue of fairness and justice among AI teachers, we can take the following approaches:Selection: Highlight all author and affiliation lines.

a) ***Ensure neutrality and diversity of sources****: when building AI teachers, we must ensure neutrality and diversity of data sources. This means that we need to collect data from a variety of perspectives and viewpoints to avoid the existence of bias and discrimination. At the same time, we should obtain data from different cultural, social, and economic backgrounds to ensure that AI teachers can understand and adapt to various learning environments and needs.*

b) ***Create AI detectors for harmful information****:* To prevent AI teachers from generating harmful information, we can utilize AI technology to create a detector for harmful information. Such detectors can automatically identify and filter information that could harm students, thus protecting them from inappropriate or harmful content.

c) ***Increase Transparency of Large Company Research***: To ensure that AI teachers are fair and equitable, we must increase the public's transparency of ample company research. Companies must disclose their research methods, data sources, and results so that the public can understand and monitor their work. This will also help increase public trust and acceptance of AI teachers.

d) ***Government enacts new laws to regulate AI teachers strictly***: To protect the rights and privacy of students, the government needs to strictly regulate the research and use of AI teachers by enacting new laws. These laws should specify how student data will be collected, used, and stored and how ethical and legal issues may be handled. At the same time, the government needs to set up specialized agencies to enforce these laws and impose severe penalties for violations. Change the number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

By taking these steps, we can better achieve fairness and justice for AI teachers, ensure that every student has equal educational opportunities, and avoid the inequalities AI technology has brought.

B. Representational Bias

Representation bias is a problem that exists in numerous domains and contexts, and it refers to the over- or under-representation of a group in a domain or context relative to its proportion in the overall population [13, 16]. This bias usually stems from deep-rooted reasons such as social structure and cultural perceptions, and in the field of AI, the existence of this bias may lead to unjust and unfair decisions and behaviors of AI systems. There are three main forms of representation bias: harmful stereotyping, underrepresentation, and overrepresentation [6]

1) Bias is Everywhere

Representation bias is a pervasive problem in AI [11]. This bias can occur in machine learning, deep learning, computer vision, or generative AI. For example, a model that improves the clarity of a picture may automatically change Obama's skin color to white due to the model being exposed to too many white faces in its training data, causing it to overrepresent white people. As another example, a model that generates photos of CEOs may generate young male CEOs because it has been exposed to an overabundance of young male CEOs in the training data, causing the model to overrepresent them [9]. [54] shows that Chat-GPT is heavily biased toward liberal politicians. All these examples show that representation bias is a real problem in AI and needs to be actively tackled.

2) Solutions

This article argues that addressing the problem of representational bias needs to be approached from several fronts. First, we need to build better communities and educate more people to understand and recognize the problem of representation bias through education, a commitment to move beyond the binary treatment of gender, more nuanced treatment of race, etc. Second, we can reduce bias in AI systems through personalized learning, allowing AI systems to personalize instruction based on each individual's characteristics and needs rather than relying solely on the dominant group in the training data[35]. Again, we need to involve experts from different backgrounds in the research process, including minorities, subcultures, people with disabilities, researchers from different countries, etc., and their participation can help us better understand and address the issue of representation bias. In addition, we can use better methods to collect data [57], build better models, and develop better norms to ensure that the decisions and behaviors of AI systems reflect the needs and interests of the majority fairly and equitably.

C. Environmental Benefits

a) Total value formula considering environmental factors: When assessing AI teachers' social impact, we must consider environmental factors. The environmental costs of large-scale AI models have attracted scholarly attention [6]. As one type of such model, the environmental impact of AI teachers is equally worthy of our consideration. Appropriating from the value formula in [6], we can argue that the social benefits of AI teachers are equal to the sum of acceptability, improved education, distance from harm, school efficiency, and environmental impact. This formula can help us assess the value of AI teachers more comprehensively than just from the perspective of educational effectiveness.

b) Unfair environment: However, it is also essential to recognize that the economic benefits and social costs of environmental problems may be unevenly distributed among communities. For example, the health impacts of establishing a toxic waste site in Silicon Valley are unevenly distributed to socially disadvantaged groups [6]. Pollution from manufacturing in Taiwan is associated with chronic health problems [6]. These are the environmental problems that AI teachers may bring. These problems make it impossible to ignore the environmental problems that AI teachers may bring about in promoting them.

c) Environmental Impact: Furthermore, environmental issues are not a far-fetched topic. There are already many environmentally related issues affecting the lives of many people. The popularity of AI teachers may contribute to this unfair consumption of resources. In addition, there will be many irreversible environmental impacts, and carbon offsets have been proposed to mitigate this. Carbon offsetting has been talked about in many papers [58,59]. Besides the fact that it does not stop people from using high-emission transportation [60], most carbon offsets are worse than solutions that do not emit CO2 in the first place [6]. Using carbon offsets can lead to more carbon emissions than if the carbon had never been emitted. This means that we need to find more effective solutions. This requires us to consider the possible environmental impacts of AI teachers as we promote them and to take adequate measures to mitigate such impacts.

Issues Facing AI Tutors' Ethics

A. Data Bias

When discussing ethical issues for AI teachers, the first thing to consider is data bias. Take OpenAI's GPT-3, for example [35], an advanced natural language processing model whose training data includes many web pages, Reddit posts, and Wikipedia as of 2021. However, these data sources are only somewhat unbiased. Reddit is a forum used mainly by young males, and statistically, more than half of its users belong to this group [1]. In contrast, only 8.8%-15% of Wikipedia's users are female [2]. They largely reflect the views and behaviors of young users in developed countries [61], which are not representative of the views and behaviors of everyone around the globe. This means that if an AI teacher relies mainly on these data for training, it may have biases in terms of gender, age, and geography when dealing with problems. This bias may lead to a lack of understanding and respect for the needs and perspectives of particular groups of people in the AI teacher's teaching process, thus violating the principles of fairness and equity in education.

B. Who will monitor?

The issue of oversight of the educational system is also an essential aspect of the ethical issues of AI teachers. Who should oversee the behavior and decision-making of AI teachers in an AI educational environment? Is it the government, schools, parents, or other stakeholders?

In a survey paper [5], while all participants expressed a desire to explore, better understand, and engage with the ethical issues surrounding the design and application of AI in educational environments, some participants seemed to believe that we, as a community, are already "doing the ethical" act because we are operating in the field of education out of the best of intentions, with the best of intentions, which is in itself ethical. This disdain for supervision is shared, either because people do not know enough about the ethical issues of AI teachers or because they do not think these issues are essential. However, this attitude can lead to the ethical issues of AI teachers being ignored, which can hurt students and the education system.

C. Be United

Finally, we must face achieving consistency in ethical rules across different education systems. The diversity of education systems makes it difficult to agree on standard rules. Different countries, regions, and schools may have different educational goals, teaching methods, and evaluation standards, which makes it very difficult to develop a common ethical rule for AI teachers that applies to all educational systems.

However, this does not mean that we should give up trying. Instead, we should strive to find common ground and create a widely accepted ethical framework for AI teachers based on shared values. This may require a great deal of research, discussion, and negotiation on our part, but only in this way can we ensure that AI teachers are applied globally with the same ethical principles, thus truly realizing justice and equity in education.

Conclusions

After we delve into the social impact of AI teachers, we can draw some conclusions and recommendations.

First, we need proactive interventions. This means we need preventive ethical risk reduction starting with model development and deployment. This may involve reviewing AI teachers' training data to ensure their sources' neutrality and diversity to minimize potential bias. In addition, we need to develop and follow several ethical guidelines and standards to ensure that AI teachers do not violate user privacy, mislead students, or negatively impact educational equity.

Second, we need to be responsive. This means we should be responsive to known errors, actively seek out possible problems, and resolve them quickly and accurately. This may require us to continually monitor and evaluate the use of AI teachers to identify and address emerging issues promptly. In addition, we will need to make changes in the future to prevent similar issues from recurring.

Then, we need to emphasize community values. We need to expand the diversity of researchers and model users to build better community values and behavioral norms. This may require us to encourage and support people from different backgrounds to participate in the research and use of AI Teacher to ensure that AI Teacher meets the needs of diverse users and respect and reflects the values of diverse users.

Finally, we need a practical regulatory framework. Governments need to update existing laws or create new ones to address the unique challenges AI-generated content poses. This could include regulations on intellectual property, plagiarism, data protection, and privacy. There is also a need to involve parents to some extent in the monitoring process to ensure that the use of AI teachers does not negatively impact the safety and health of their children.

Overall, AI teachers have enormous potential to bring many benefits to education. However, we must also realize that using AI teachers may bring ethical and social issues. Therefore, we need to take proactive measures to ensure that AI teachers can benefit education rather than problems.

References

[1] Bloom, B.S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, [online] 13(6), pp.4–16. Doi https://doi.org/10.2307/1175554.

[2] openai.com. (n.d.). *Khan Academy*. [online] Available at: https://openai.com/customer-stories/khan-academy.

[3] Heaven, W.D. (2023). *ChatGPT is going to change education, not destroy it*. [online] MIT Technology Review. Available at: https://www.technologyreview.com/2023/04/06/1071059/chatgpt-change-not-destroy-education-openai/.

[4] Chen, C. (2023). *AI Will Transform Teaching and Learning. Let's Get it Right.* [online] Stanford HAI. Available at: https://hai.stanford.edu/news/ai-will-transform-teaching-and-learning-lets-get-it-right.

[5] Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S.B., Santos, O.C., Rodrigo, M.T., Cukurova, M., Bittencourt, I.I. and Koedinger, K.R. (2021). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*, 32, pp.504–526. Doi https://doi.org/10.1007/s40593-021-00239-1.

[6] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D. and Donahue, C. (2021). On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]*. [online] Available at: https://arxiv.org/abs/2108.07258.

[7] Poesia, G. and Goodman, N.D. (2023). Peano: Learning Formal Mathematical Reasoning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, [online] 381(2251), p.20220044. doi https://doi.org/10.1098/rsta.2022.0044.

[8] Anthony, T., Tian, Z. and Barber, D. (2017). *Thinking Fast and Slow with Deep Learning and Tree Search*. [online] arXiv.org. Available at: https://arxiv.org/abs/1705.08439.

[9] Adrien Book. *What Can AI-Generated Images Tell Us About Society's Biases?* Available at: https://www.wearedevelopers.com/magazine/generative-ai-image-bias [Accessed 17 Jul. 2023].

[10] https://www.sohu.com/a/18373609_105067

[11] Nicoletti, L. and Bass, D. (2023). Humans Are Biased. Generative AI Is Even Worse. *Bloomberg.com*. [online] Available at: https://www.bloomberg.com/graphics/2023-generative-ai-bias/?utm_medium=deeplink&leadSource=uverify%20wall [Accessed 17 Jul. 2023].

[12] Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y. and Pfister, T. (2023). *Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes*. [online] arXiv.org. Doi:https://doi.org/10.48550/arXiv.2305.02301.

[13] Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N.A. (2019). The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Doi https://doi.org/10.18653/v1/p19-1163.

[14] Vidgen, B. and Derczynski, L. (2020). Directions in Abusive Language Training Data: Garbage In, Garbage Out. *arXiv:2004.01670 [cs]*. [online] Available at: https://arxiv.org/abs/2004.01670.

[15] Weng, L. (2021). *Reducing Toxicity in Language Models*. [online] lilianweng.github.io. Available at: https://lilianweng.github.io/posts/2021-03-21-lm-toxicity/ [Accessed 17 Jul. 2023].

[16] Zhou, K., Ethayarajh, K. and Jurafsky, D. (2021). Frequency-based Distortions in Contextualized Word Embeddings. *arXiv:2104.08465 [cs]*. [online] Available at: https://arxiv.org/abs/2104.08465.

[17] Guo, C. (2022). Research on Improvement of College Teachers' Teaching Abilities in the Artificial Intelligence Era. International Journal of Scientific Advances (IJSCIA), Volume 3| Issue 4: Jul-Aug 2022, Pages 581-583, URL: https://www.ijscia.com/wp-content/uploads/2022/08/Volume3-Issue4-Jul-Aug-No.311-581-583.pdf

[18] Dimitriadis, G. (2020). Evolution in Education: Chatbots. *Homo Virtualis*, 3(1), p.47. Doi https://doi.org/10.12681/homvir.23456. [19] Velander, J., Taiye, M.A., Otero, N. and Milrad, M. (2023). Artificial Intelligence in K12 Education: eliciting and reflecting on Swedish teachers' understanding of AI and its implications for teaching & learning. *Education and Information Technologies*. [online] doi https://doi.org/10.1007/s10639023119904.

[19] Haruna-Cooper, L. and Rashid, A. (2023). GPT-4: the future of artificial intelligence in medical school assessments. *Journal of the Royal Society of Medicine*, 116(6). Doi https://doi.org/10.1177/01410768231181251.

[20] Zhang, S.J., Florin, S., Lee, A.N., Niknafs, E., Marginean, A., Wang, A., Tyser, K., Chin, Z., Hicke, Y., Singh, N., Udell, M., Kim, Y., Buonassisi, T., Solar-Lezama, A. and Drori, I. (2023). *Exploring the MIT Mathematics and EECS Curriculum Using Large Language Models*. [online] arXiv.org. Doi https://doi.org/10.48550/arXiv.2306.08997.

[21] Duolingo Team (2023). *Introducing Duolingo Max, a learning experience powered by GPT-4*. [online] Duolingo Blog. Available at: https://blog.duolingo.com/duolingo-max/.

[22] Mearian, L. (2023). *Schools look to ban ChatGPT; students use it anyway*. [online] Computerworld. Available at: https://www.computerworld.com/article/3694195/schools-look-to-ban-chatgpt-students-use-it-anyway.html.

[23] Court, M. (2023). *Why top UK universities will incorporate 'ethical' use of Chat GPT in education*. [online] The Northern Echo. Available at: https://www.thenorthernecho.co.uk/news/23631319.chat-gpt-uk-universities-incorporate-ethical-use-ai/ [Accessed 17 Jul. 2023].

[24] Chesney, R., & Citron, D. K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. California Law Review, 107, 1753–1819.

[25] Rini, R. (2020). Deepfakes and the Epistemic Backstop. Philosophers' Imprint, 20(9), 1-21.

[26] Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. Science, Technology, & Human Values, 41(1), 118–132.

[27] Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37–43.

[28] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951720915687.

[29] Owain Evans. How truthful is GPT-3? A benchmark for language models. *www.alignmentforum.org*. Available at: https://www.alignmentforum.org/posts/PF58wEdztZFX2dSue/how-truthful-is-gpt-3-a-benchmark-for-language-models.

[30] Sobieszek, A. and Price, T. (2022). Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models. *Minds and Machines*. Doi https://doi.org/10.1007/s11023-022-09602-0.

[31] Dey, V. (2021). Are Larger Language Models Less Truthful? *Analytics India Magazine*. Available at: https://analyticsindiamag.com/are-larger-language-models-less-truthful/ [Accessed 17 Jul. 2023].

[32] Gordon, R. (2023). 3 Questions: Jacob Andreas on large language models. *MIT News | Massachusetts Institute of Technology*. Available at: https://news.mit.edu/2023/3-questions-jacob-andreas-large-language-models-0511.

[33] Wiggers, K. (2021). *Falsehoods are more likely with large language models*. [online] VentureBeat. Available at: https://venturebeat.com/business/falsehoods-more-likely-with-large-language-models/.

[34] Johnson, S. and Iziev, N. (2022). A.I. Is Mastering Language. Should We Trust What It Says? *The New York Times*. [online] 15 Apr. Available at: https://www.nytimes.com/2022/04/15/magazine/ai-language.html.

[35] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C. and Hesse, C. (2020). Language Models are Few-Shot Learners. *arxiv.org*. [online] Available at: https://arxiv.org/abs/2005.14165.

[36] Winn, Z. (2021). *Enabling AI-driven health advances without sacrificing patient privacy*. [online] MIT News | Massachusetts Institute of Technology. Available at: https://news.mit.edu/2021/secure-ai-labs-health-care-1007.

[37] The NSW Department of Education (2020). *Teaching privacy and ethical guardrails for the AI imperative in education – Public Interest Privacy Center*. [online] Public Interest Privacy Center. Available at: https://publicinterestprivacy.org/teaching-privacy-and-ethical-guardrails-for-the-ai-imperative-in-education/ [Accessed 17 Jul. 2023].

[38] Hosseini, M. and Horbach, S.P.J.M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for using ChatGPT and other Large Language Models in scholarly peer review.

Research Square, [online] pp.rs.3.rs2587766. Doi:https://doi.org/10.21203/rs.3.rs-2587766/v1.

[39] Perttu Hämäläinen, Mikke Tavast and Kunnari, A. (2023). Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. Doi:https://doi.org/10.1145/3544548.3580688.

[40] Al-Rubaie, M. and Chang, J.M. (2019). Privacy-Preserving Machine Learning: Threats and Solutions. *IEEE Security & Privacy*, 17(2), pp.49–58. doi:https://doi.org/10.1109/msec.2018.2888775.

[41] Carlini, N. (2020). *Privacy Considerations in Large Language Models*. [online] ai.googleblog.com. Available at: https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html?m=1 [Accessed 17 Jul. 2023].

[42] Dinan, E., Humeau, S., Chintagunta, B. and Weston, J. (2019). *Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.1908.06083.

[43] www.perspectiveapi.com. (n.d.). *Perspective API - How it works*. [online] Available at: https://www.perspectiveapi.com/how-it-works/.

[44] Gongane, V.U., Munot, M.V. and Anuse, A.D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12(1). doi:https://doi.org/10.1007/s13278-022-00951-3.

[45] Meel, P. and Vishwakarma, D.K. (2021). A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles. *Expert Systems with Applications*, 177, p.115002. doi:https://doi.org/10.1016/j.eswa.2021.115002.

[46] Guacho, G.B., Abdali, S., Shah, N. and Papalexakis, E.E. (2018). *Semi-supervised Content-based Detection of Misinformation via Tensor Embeddings*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.1804.09088.

[47] Bayer, A. and Wilcox, D.W. (2019). The unequal distribution of economic education: A report on the race, ethnicity, and gender of economics majors at U.S. colleges and universities. *The Journal of Economic Education*, 50(3), pp.299–320. doi:https://doi.org/10.1080/00220485.2019.1618766.

[48] Hylen, J. (2006). *(PDF) Open educational resources: Opportunities and challenges*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/235984502_Open_educational_resources_Opportunities_and_challenges.

[49] RICH, M., COX , A. and BLOCH, M. (2016). Money, Race and Success: How Your School District Compares. *cepa.stanford.edu*. [online] Available at: https://cepa.stanford.edu/news/money-race-and-success-how-your-school-district-compares [Accessed 17 Jul. 2023].

[50] Gómez, E., Shui Zhang, C., Boratto, L., Salamó, M. and Marras, M. (2021). The Winner Takes it All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. doi:https://doi.org/10.1145/3404835.3463235.

[51] Peng, M., Xie, J., Xiong, M. and Liu, Y. (2023). Artificial Intelligence Education in Primary and Secondary Schools from the Perspective of Thinking Quality. *Journal of Contemporary Educational Research*, 7(4), pp.41–46. doi:https://doi.org/10.26689/jcer.v7i4.4875.

[52] Mahmud, B.U., Hong, G.Y. and Fong, B. (2022). A Study of Human-AI Symbiosis for Creative Work: Recent Developments and Future Directions in Deep Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications*. doi:https://doi.org/10.1145/3542698.

[53] Brynjolfsson, E. (2022). *The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence*. [online] Stanford Digital Economy Lab. Available at: https://digitaleconomy.stanford.edu/news/the-turing-trap-the-promise-peril-of-human-like-artificial-intelligence/.

[54] McGee, R.W. (2023). *Capitalism, Socialism and ChatGPT*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4369953.

[55] Montes, G.A. and Goertzel, B. (2019). Distributed, decentralized, and democratized artificial intelligence. *Technological Forecasting and Social Change*, [online] 141, pp.354–358. doi:https://doi.org/10.1016/j.techfore.2018.11.010.

[56] Chen, Y., Mahoney, C., Grasso, I., Wali, E., Matthews, A., Middleton, T., Njie, M. and Matthews, J. (2021). Gender Bias and Under-Representation in Natural Language Processing Across Human Languages. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. [online] doi:https://doi.org/10.1145/3461702.3462530.

[57] Li, Y. and Vasconcelos, N. (2019). *REPAIR: Removing Representation Bias by Dataset Resampling*. [online] openaccess.thecvf.com. Available at: https://openaccess.thecvf.com/content_CVPR_2019/html/Li_REPAIR_Removing_Representation_Bias_by_Dataset_Resampling_CVPR_2019_paper.html.

[58] Sapkota, Y. and White, J.R. (2020). Carbon offset market methodologies applicable for coastal wetland restoration and conservation in the United States: A review. *Science of The Total Environment*, 701, p.134497. doi:https://doi.org/10.1016/j.scitotenv.2019.134497.

[59] Peters, L.J.B., Chattopadhyay, G. and Kandra, H.S. (2021). *Carbon offsetting in the road transport industry: issues and challenges of meeting the objectives*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICMIAM54662.2021.9715213.

[60] Bösehans, G., Bolderdijk, J.W. and Wan, J. (2020). Pay more, fly more? Examining the potential guilt-reducing and flight-encouraging effect of an integrated carbon offset. *Journal of Environmental Psychology*, 71, p.101469. doi:https://doi.org/10.1016/j.jenvp.2020.101469.

[61] Lucy, L. and Bamman, D. (2021). *Gender and Representation Bias in GPT-3 Generated Stories*. [online] ACLWeb. doi:https://doi.org/10.18653/v1/2021.nuse-1.5.