IJHSR
ijhighschoolresearch.org
REVIEW ARTICLE

The Morality of Large Language Models

Ronan Cheng-Slater

Hopkins School, 986 Forest Road, New Haven, CT 06515, U.S.A ; chengslaterronan@gmail.com

ABSTRACT: We live in a time when artificial intelligence has taken the world by storm. Amidst this storm, one chatbot, ChatGPT (Chat Generative Pre-trained Transformer), has captured the public's attention like no other. Before ChatGPT most of the public did not know or have reason to care about artificial intelligence in the form of large language models (LLM), but when OpenAI released ChatGPT-3 to the public, everything changed. Unlike previous chatbots, ChatGPT-3 was more accessible to the public and was uncannily good at mimicking human conversation and writing. Since ChatGPT-3 can replicate human conversation, answer questions, and summarize information, society must now grapple with the ethics involved in the making and the use of ChatGPT, such as intellectual property theft stemming from the use of copyrighted works to train ChatGPT, bias arising out of flawed training data sets, privacy concerns from potential data leaks, the spread of false or misleading information, transparency, and accountability. This paper reviews the ethical challenges in making and using ChatGPT and how the creators of LLMs are trying to overcome such challenges.

KEYWORDS: Robotics and Intelligent Machines; Machine Learning; Large Language Models; ChatGPT.

Introduction

Generative AI is artificial intelligence technology capable of generating text, images, code, or other types of content through machine learning systems. Utilizing a computing process known as deep learning, generative AI analyzes patterns in large sets of data and then replicates the patterns to create new data to mimic how a human would respond.[1] Unlike humans, however, generative AI does not possess an inherent sense of right and wrong. This paper aims to explore the ethical issues arising from the creation and use of generative AI and does so by examining the latest generation of generative AI, ChatGPT. To fully appreciate the ethical problems with ChatGPT, the paper first reviews the history of its development, then considers how it works at a more technical level regarding how it works, and then examines the various ethical problems that arise from how the models are trained and used and possible ways to address these issues.

Background

Brief History of Generative AI:

Generative AI has existed for many more years than most people realize. Dating back to the 1950s, generative AI came in the form of hidden Markov and Gaussian mixture models.[2] Those models could not do nearly as much as today's generative AI models as they could only process sequential data such as speech and time series. The recent development of deep learning has caused generative AI to become much more advanced.[3,4] Early generative AI that did natural language processing (NLP) used N-gram language modeling to learn word distribution and predicted the next word through the best fit to create sentences.[5] While this worked for short sentences, N-gram language modeling proved less workable for longer sentences as the model became increasingly less accurate with each additional word added to the sentence. To mitigate the length issue, computer scientists developed recurrent neural networks (RNNs), which were used for language modeling tasks.[6] Shortly after that another advance in generative AI came in the form of the development of long short-term memory (LSTM). LSTM was developed, which allowed the generative AI model to have its short-term memory last a longer number of timesteps, thus increasing its accuracy.[7] Another significant step was the recent development of deep learning which has caused generative AI to become much more advanced.[3,4]

Today, one of the chatbots at the forefront of the deep learning frontier is ChatGPT. When OpenAI released ChatGPT-3 on November 30, 2022, the model received enormous attention from the public and experts alike over its preternatural ability to imitate human text and answer complex questions with accuracy far surpassing that of older chatbots.[8] Many major companies and entrepreneurs have since been looking to invest in this new wave of technology and some have even been trying to rush out their own AI chatbots, such as Google Bard and Microsoft's Azure.

Inner Workings of ChatGPT:

ChatGPT is a large language model (LLM). An LLM is a type of AI that uses neural networks (deep learning) and large data sets to comprehend, condense, generate, and predict content.[8, 9] There are two main phases in the process of creating ChatGPT, (i) data gathering and training, and (ii) user interaction. The first phase, data gathering and training, is where most of ChatGPT is created. In this phase, two types of training are employed: supervised and unsupervised. Supervised training is where inputs with set outputs are used to train the chatbot, whereas unsupervised training is when inputs are used without any set output.[10] Unsupervised training helps ChatGPT increase the breadth of answers while supervised training is used to help fine-tune ChatGPT for issues such as toxicity and

DOI: 10.36838/v6i7.15

bias.[11] An input transitions to an output through something called neural networks. Neural networks are made of input layers, output layers, hidden layers, and the node that connects them. Some neural networks can have multiple hidden layers such as the neural network in Figure 1.[8]

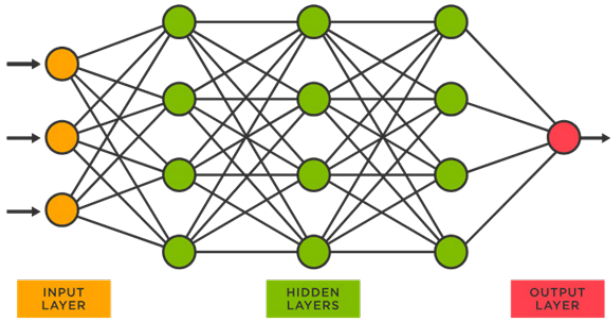

Figure 1: A visual representation of a neural network.[8]

What each layer does is as follows: the input layer is where the data is first entered into the neural network, the hidden layer is where the information is processed, and finally the output layer is where the system decides how to proceed. Most animal brain structures, including our own, inspire this basic structure. Both intake information, process it and learn from it. The material difference is that the information humans and animals take in is from our senses and not a data set.[8]

The system understands words through tokenization. Tokenization is the breakdown of sentences into individual tokens, which then are embedded, allowing the computer to understand the human text, similar to how humans have to translate one language to another to understand each other.[12] Figure 2 is a visual representation of tokenization.

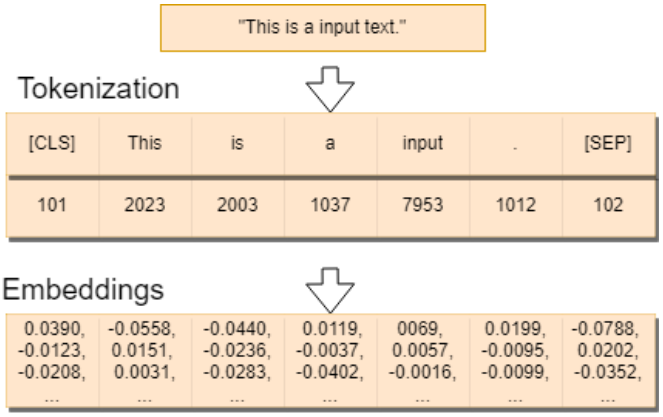

Figure 2: A visual representation of tokenization.[12]

Embedding is when an LLM like ChatGPT takes the individual words made during tokenization and converts them into number vectors the computer understands.

Ultimately, though ChatGPT does not process language as humans do, it can understand language syntax and semantics by converting words into numbers and reviewing large datasets of text.[11] ChatGPT then predicts the next most likely word to simulate human writing. ChatGPT's ability to replicate human with massive data sets. The free version of ChatGPT-3 has 175 billion parameters used to learn and understand the information given to it. Furthermore, companies like Microsoft are optimizing the storage and efficiency of supercomputers, which, in turn, allow chatbots to have more parameters, leading to even more realistic human-like writing.[10] Due to ChatGPT drawing from these large data sets from the internet, it is susceptible to hallucinations. Hallucinations are when ChatGPT generates something with a degree of falsity. This can happen because some of the data sets ChatGPT used had false information or errors in encoding and decoding. The false information ChatGPT generates can lead to unintentional injuries and spreading misinformation.

Given the power of ChatGPT and its progeny, society must inevitably turn to the ethical implications of creating and using this evolving technology.[13-15] The subsequent section reviews ethical considerations in the making of an LLM, specifically touching upon concerns of intellectual property theft, possible improper use of publicly and quasi-publicly available data in training the model and finally privacy risks to those whose data has been used in the training process. The paper will then consider hidden and latent issues with how LLMs formulate their output, including bias from the source materials used to train the model. The paper will also consider challenges to ethical norms in using LLMs such as intellectual property rights, privacy, transparency, and accountability. Finally, we will conclude with a discussion of proposals to overcome such challenges.

Ethical Concerns Arising from ChatGPT

Intellectual Property Theft:

To be effective, generative AI, including LLMs, needs vast quantities of data upon which to be trained. Data for such purposes comes from various places, some protected by intellectual property laws. The ethics of using copyrighted materials without the writers' permission has raised some questions and resulted in some lawsuits being filed against the AI companies for scrapping the copyrighted, protected works for training the AI companies' models. For example, a class-action lawsuit was recently filed by three authors in federal court in California against OpenAI, seeking an injunction against the continuing use of the writers' copywritten material and financial damages. The plaintiffs allege that OpenAI trained its AI LLM by copying the text of the authors' work and extracting expressive information from it.[16] The complaint notes that "much of the material in OpenAI's training datasets…comes from copyrighted works - including books written by Plaintiffs - that were copied by OpenAI without consent, without credit, and …compensation." The plaintiffs also filed a lawsuit against Meta alleging similar copyright infringement claims.[16]

The intellectual property question is currently under debate in the US court system as artists, writers and publishers sue OpenAI and other operators. For example, the New York Times sued OpenAI and Microsoft for "unlawful use of The Times's work to create artificial intelligence products that compete with it" and "threatens The Times's ability to provide that service."[17] The lawsuit provides examples of ChatGPT

answers that are almost verbatim from New York Times articles and unattributed to its source.[17]

While the case for an ethical breach by AI companies using copyrighted materials without permission may seem obvious, a more thorough analysis is warranted. Consider, for example, the case of the fair use exception to copyright law. According to the U.S. Copyright Office, fair use allows for the limited use of copyrighted material without permission for certain purposes including reporting, teaching, scholarship, or research.[18] Training LLMs could fall under the umbrella of fair use, especially if the use is transformative and does not substitute for the original work. As noted above, LLMs do not reproduce the original copyrighted works word for word but rather extract statistical patterns from the data to generate new text. It is thus transformative and, as such, within the general fair use exception to copyright protection. In the case of the New York Times, the extent of verbatim use of source articles argues against the Fair Use doctrine requirement that the output be "transformative use."[18]

However, proponents of the free use of copyrighted materials for training AI models argue that the purpose of LLM is not to store copyrightable content but to train algorithms on language. Andreessen Horowitz (a16z) submitted a comment in response to the U.S. Copyright Office's August 30 Notice of Inquiry on Artificial Intelligence and Copyright and stated that generative AI model training is a "productive, non-exploitive use of training material." A16z further argues that generative AI requires "an almost unimaginably massive amount of content."[19] They further argue that without massive amounts of diverse content, generative AI models will be "unable to accurately discern the meanings of or semantic relationships among words as used in human language broadly."[19] Thus, the cost, even if nominal per copyrighted piece, and the regulatory framework required to enforce this would have a chilling effect on the US's development of AI models. The evidence in the New York Times lawsuit of verbatim copying of articles is troubling. Still, it should be viewed as an exception that must be corrected in the model rather than emblematic of a core issue with generative AI and its use of copyrighted material.[20]

As copyrighted texts contain more information and knowledge than in the public domain, allowing LLMs access to such works will likely provide better outputs for educational and research purposes. This, in turn, may spur greater innovation and learning and thus benefit society in the long run.

Privacy Breaches:

Privacy concerns arising from LLMs come in two flavors: using people's data without their consent and the potential for data leaks. As noted previously, LLMs are trained on data that typically is scraped from the internet. In most cases, individuals sharing such information online do not give explicit consent for their data to be used in training AI models. This raises the ethical question of whether such individuals have a right to control how their data is used when shared in public or quasi-public forums. In the case of quasi-public forums, for example, social media sites, most users/creators would likely have some reason to believe that their data was not fully private but may not have shared their information had they known the use to which LLMs would put it. Without a clear agreement between the creator of the content and the social media site, there appears to be a gap between reasonable expectations and the reality of how LLMS acquires its data. The argument favoring an ethical breach is more tenuous in the case of data published generally on the internet. It would seem that the authors of such works should have expected no ongoing control over their data. Interestingly, the European Union has considered this question and determined that ChatGPT and other chatbots use people's data for their training sets without consent, which is illegal under the EU's General Data Protection Regulation (GDPR).[21] Some countries, such as Italy, have banned ChatGPT due to privacy considerations.[22] To date, United States regulations are much more limited. The other privacy issue, as noted above, relates to the possibility of data leakages, where an assailant tries to extract data from the chatbot leading to people's data being stolen.[23] LLMs present rich targets given the data needed to train the model. The question for society is whether it wants to allow a private company to control that kind of data without the explicit consent of the persons whose data is at risk.

Bias:

One might think that a bot, being a machine, would be incapable of being biased, sexist, or racist, but because of how ChatGPT is trained, the seemingly harmless chatbot can become a source of harmful content. During ChatGPT's unsupervised training, it can accidentally pick up harmful material and be trained to propagate such things.[24] Real-world examples of this already exist. Twitter (or X as of 2023) released their interactive chatbot, Tay, designed to use Twitter as its data set to learn how to talk like a teenager. Because of the Twitter universe's toxic data set, Tay quickly became a racist neo-nazi. Programmers at Twitter did not predict that the data that they drew from could contain toxicity. Thus, they failed to anticipate Tay's significant bias in responding to user input.[25] While Twitter's Tay was a more concentrated and extreme example of the internet affecting a chatbot, the same concept applies to all chatbots, including ChatGPT. Machine learning has also been found to discriminate based on classes like race and gender. For example, a study by Tolga Bolukbasi *et al*. found biases in machine learning-based output, associating the word "woman" with the word "homemaker."[26] ProPublica found a criminal justice algorithm that discriminated against blacks. The program scored people who were arrested by their likelihood of committing a future crime and consistently scored blacks at higher risk than whites regardless of background and criminal record.[27] A study by the Brookings Institution also found that ChatGPT had a strong left-leaning bias.[28]

Lack of Accountability for Harms Produced by ChatGPT:

A final issue posed by this new technology concerns accountability. If ChatGPT develops an incorrect answer, whose fault is it? As noted before, chatbots are susceptible to hallucination. One can readily imagine how a hallucination in the wrong circumstance may result in significant harm in the real world if the information produced by the chatbot is wrong on a key point. For example, the chatbot may give a user with faulty information about proper medicine dosages or the appropriate

DOI: 10.36838/v6i7.15

way to treat an injury. Should the developers be able to avoid liability if their software comes with a blanket disclaimer? Should the chatbot always highlight the possibility that its output is wrong and that the user should beware? Is that enough? The more society comes to rely upon this new technology, the greater the stakes for users and developers alike, and the more likely it becomes that conflicts over liability will end up seeking resolution in our court system.

Furthermore, Elon Musk believes that the development of ChatGPT and generative AI models should be done for something other than profit. Musk argues that the importance of generative AI and the work of OpenAI is extraordinarily impactful and embedded with risk and should not be skewed by the goal of personal profit of an individual or company. He believes that OpenAI should be open-source and non-profit. Elon Musk has recently brought the issue of accountability to light by filing suit against OpenAI for, amongst other things, breach of fiduciary duty and unfair competition.[29] Musk's concern centers even more broadly over the unregulated and uncontrolled development of generative AI to "achieve superhuman machine intelligence" that would pose "the greatest threat to the continued existence of humanity." OpenAI was established "for the good of the world" in which "safety should be a first-class requirement." It was deliberately structured as a non-profit to "collaborate widely and shift the dialog towards being about humanity winning rather than any particular group or company."[29]

Regardless of the extent of harm one may believe ChatGPT may pose - from harmful information to the replacement of humans - it remains important to consider who should be held accountable for damage resulting from the output of ChatGPT and other LLMs.

Discussion

The issue of intellectual property rights about LLM's usage of copyrighted material has only just begun being explored. The various lawsuits that have been filed center around the concept of Fair Use. While the New York Times has found errors and omissions in the output of ChatGPT in their lawsuit, the basic concept of the LLM arguably falls squarely within the definition of the transformation of information required by the Fair Use doctrine. Furthermore, ChatGPT and other LLMs have reached a tipping point of creating output that has proven useful and fascinating to the public. As a16z stated, "Generative AI promises tremendous societal benefits: from empowering the disabled to delivering world-class educational resources to underserved communities to solving some of humanity's most pressing and intractable scientific and medical problems. Allowed to reach its full potential, AI can be a tool for enhancing–not undermining–human innovation and creativity." Proposed solutions of compensation and prohibition will likely have a chilling effect on the continued development of AI in the United States and put the United States at a disadvantage in what could arguably be the most important economic development for this decade. However, creating more transparency regarding the LLM process will allow the sources to be traced. This, in turn, will allow outputs to be credited to sources if necessary.

The issue of bias results from bias in the pool of source data and bias in the training model. ChatGPT and LLMs need a pool of data that is as diverse and massive as possible for LLM training, partially to address this issue of bias. In addition, to guard against bias, OpenAI, the creators of ChatGPT, implemented something they call guardrails to keep the chatbot from being biased, sexist, racist, etc.[23] These guardrails act as a filter where when a person would ask a controversial question, ChatGPT would respond with a pre-programmed response telling the user how the chatbot could not answer such questions. Although OpenAI added these guardrails, users have found, intentionally or not, ways to get around these safeguards through so-called "jailbreaking."[23] Jailbreaking happens when someone, usually the user, tries to bypass the guidelines or guardrails in place. Although jailbreaking OpenAI's ChatGPT-3 is harder than jailbreaking its earlier models, ChatGPT-1 and ChatGPT-2, OpenAI continues to work on and strengthening ChatGPT's guardrails through a variety of methods, including trial and error.[30] Since the developers cannot anticipate all the ways ChatGPT can be used, their work will continue into the foreseeable future.[31]

In addition to choosing what questions to answer and what data to use, OpenAI uses human feedback to train its AI models.[31] Developers are attempting to reduce the reliance on human training and create more automated aspects of the training with embedded moral training. For example, Dan Hendrycks *et al.* created an "environment suite of 25 text-based adventure games with thousands of diverse, morally salient scenarios." The games attempt to reward moral actions.[32] While the issue of morality can become highly philosophical, these training mechanisms should continue to be developed and OpenAI and other developers should continue to create moral training for its LLMs. As well there is also the emerging field of representation engineering, which uses concepts from cognitive neuroscience to assist in manipulating the machine's "cognitive phenomena" to address issues of bias, honesty, hallucinations, harmlessness, and power-seeking.[33] While bias is perhaps impossible to eliminate, owners of AI must continue to develop mechanisms and training, whether human or machine, to reduce bias. Users must also recognize that bias will exist, no matter how advanced the LLM is. Understanding how to read and analyze output should become an important part of user education as ChatGPT and LLMs become more prevalent and commonly used.

Lastly, the issue of accountability is critical. As discussed above, LLMs comprise inputs, training, and outputs. Most importantly, LLMs in the United States are privately owned and run. Accountability must be established to make solutions viable, whether enforcing laws and regulations or creating machine and human training to address misinformation, bias, hallucinations, privacy and honesty. Should OpenAI (and other LLM owners) argue that there are no copyright issues through the transformative nature of the machine, then the argument would apply that OpenAI is accountable for the outputs of the LLM and any damages that they may cause. Given LLMs' complexity, a simple disclaimer should be considered insufficient to protect OpenAI and its brethren from liability.

According to a16z, "AI represents a new computing paradigm that will transform information technology and computing as fundamentally as the development of the microchip and the rise of the internet did over the past 70 years."19 Billions of dollars have been invested in the development of AI technologies.[19] Microsoft has invested $13 billion in OpenAI alone, which should be sufficient to include investing in technologies that will address the risk of bias, misinformation, hallucinations, and potential impacts to end users.[17]

One possible way forward is to begin to develop legislation that allocates responsibility between users and developers in a fair way that encourages technological innovation while at the same time protecting users from unscrupulous developers. The more that accountability is assigned, the more likely the developer and user will take greater care in considering the safety of others and the potential uses. A recent symposium at Bowdoin College suggested that creators, in addition to thinking about how ChatGPT works, consider "why" and "what if." The questions a creator can consider include: for whom is it being created (both known and unknown); who benefits; who is it exploiting (intentionally or not); and what are the risks.[13] Imposing this ethical discipline may prevent some of the potential hazards we may experience from ChatGPT.

Conclusion

In closing, AI has come far, dating back to the 1950s when AI was only a notion of a robot, and today, AI chatbots such as ChatGPT are replicating very real human text. ChatGPT and large language model improvements are now accelerating at a rapid pace through the use of generative AI. Generative AI uses tokenization to assist the machine in learning human syntax and semantics. ChatGPT and LLMs represent a new computing paradigm that will transform information technology and our economy. However, with all this progress it is important to think about the repercussions and ethics of AI at an equal pace, if not faster. There are serious issues, including intellectual property theft, privacy breaches, bias, and who should be held accountable for these things. To date, OpenAI has taken some steps to help mitigate the effects of these ethical issues, such as bias, transparency, privacy, and accountability. However, the solutions still need to address the issues fully. These rules should firmly assign accountability for damages further encouraging various participants to consider risks and safety. As Elon Musk's lawsuit indicates, economics can impact behavior; this can hold not only profits but fines and punishments as well. However, as a16z shared, whatever laws and regulations are put in place must carefully consider whether they would ultimately have such a chilling effect on the development of AI, that it puts the United States at disadvantage in the race to advance technology and the chance to shape the future. We are in the early stages of a revolution, and much research needs to focus on addressing these ethical issues in constructing large language models and in the regulatory framework that optimizes safety without hindering progress.

Acknowledgment

I would like to thank my family for their supporting my passion for computing and debating the ethics of AI. I also thank Dr. Salinna Abdullah for her guidance and feedback while drafting this paper.

References

1. Kerner, S. M. Large language model (LLM), 2023. TechTarget. https://www.techtarget.com/whatis/definition/large-language-model-LLM (Accessed: August 14, 2023).
2. Turing, A.M. (1950) I.-Computing machinery and intelligence, OUP Academic. Available at: https://academic.oup.com/mind/article/LIX/236/433/986238 (Accessed: March 1, 2024).
3. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," arXiv, 2020.
4. B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3045–3059, 2021.
5. Schick, T; Schütze, H. *Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference*. arXiv:2001.07676v2, pp. 1-10: 2021.
6. Zhang, Z; Zhang, A; Li, M; Smola, A. *Automatic Chain of Thought Prompting in Large Language Models.* arXiv:2210.03493v1, pp.1-9: 2023.
7. J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, 2022.
8. What is a Neural Network?, 2023. TIBCO Software. https://www.tibco.com/reference-center/what-is-a-neural-network#:~:text=An%20artificial%20neural%20network%20in,where%20the%20information%20gets%20processed (Accessed: August 20, 2023).
9. Kerner, S. M. Large language model (LLM), 2023. TechTarget. https://www.techtarget.com/whatis/definition/large-language-model-LLM (Accessed: August 14, 2023).
10. Gewirtz, D; Windsor, A. How does ChatGPT actually work, *ZDNET*, Sept 20, 2023, p. 1.
11. Ruby, M. How ChatGPT Works: The Model Behind The Bot, *Towards Data Science*, January 30, 2023, Retrieved from https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286.
12. Kosar, V. Tokenization in Machine Learning Explained, 2022. Vacklav Kosar. https://vaclavkosar.com/ml/Tokenization-in-Machine-Learning-Explained#:~:text=Tokenization%20is%20splitting%20the%20input,embedded)%20into%20a%20vector%20space (accessed August 18, 2023).
13. Porter, T. Considering the "Why" and Not Just the "How" of Computer Science, 2021. Bowdoin College Website. https://www.bowdoin.edu/news/2021/12/considering-the-why-and-not-just-the-how-of-computer-science.html (Accessed: Aug 22, 2023).
14. Krugel, S; Ostermaier, A; Uhl, M. *The moral authority of chatGPT*. arxiv.org/abs/2301.07098v1, pp.1-10: 2023.
15. Chomsky, N; Roberts, I; Jeffrey Watumull. Noam Chomsky: The False Promise of ChatGPT, *The New York Times*, March 8, 2023, Retrieved from https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.
16. Sarah Silverman, Christopher Golden, Richard Kadrey, Plaintiffs v. OpenAI, Inc., OpenAI, LP, OpenAI Opco, LLC, OpenAI GP, LLC, OpenAI Startup Fund GP I, LLC, OpenAI Startup Fund I, LP, OpenAI Startup Fund Management, LLC, Defendants, Case 3:23-cv-03223-AMO, July 7, 2023. https://www.documentcloud.org/documents/23869693-silverman-openai-complaint (Accessed: March 1, 2024).
17. Vincent, J. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*, May 24, 2016.

DOI: 10.36838/v6i7.15

18. Office, U.S.C. (2023) U.S. Copyright Office Fair Use Index, U.S. Copyright Office Fair use index. Available at: https://www.copyright.gov/fair-use/ (Accessed: March 1, 2024).
19. Horowitz, A. (2023) Notice of Inquiry on Artificial Intelligence & Copyright (Dkt. 2023–6), Comments of a16z October 30, 2023, Regulations.gov. Available at: https://www.regulations.gov/comment/COLC-2023-0006-9057 (Accessed: March 1, 2024).
20. The New York Times Company, Plaintiff, v. Microsoft Corporation, OpenAI, Inc., OpenAI LP, OpenAI GP LLC, OpenAI LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC, Case 1:23-cv-11195, December 26, 2023. *https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf* (accessed March 1, 2024).
21. The general data protection regulation, 2018. European Council and Council of the European Union. https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/#:~:text=The%20GDPR%20lists%20the%20rights,his%20or%20her%20personal%20data (Accessed: August 30, 2023).
22. Intelligenza artificiale: il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell'età dei minori, 2023. Garante per La Protezione Dei Dati Personali. https://www.garanteprivacy.it/web/garante-privacy-en/main-decisions (Accessed: August 30, 2023).
23. Zhuo, T. Y; Huang, Y; Chen, C; & Xing, Z. *Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity*. arXiv:2301.12867, pp. 1-11: 2023.
24. Ray, P. P. *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope.* Internet of Things and Cyber-Physical Systems. Internet of Things and Cyber-Physical Systems 3, 121-154: 2023.
25. Blogs, M.C. (2023) Microsoft and OpenAI Extend Partnership, The Official Microsoft Blog. Available at: https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/ (Accessed: March 1, 2024).
26. Zou, A. *et al*. (2023) *Papers with code - representation engineering: A top-down approach to AI transparency, Representation Engineering: A Top-Down Approach to AI Transparency | Papers With Code*. Available at: https://paperswithcode.com/paper/representation-engineering-a-top-down (Accessed: March 1, 2024).
27. Angwin, J. *et al*. (2016) Machine bias, ProPublica. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed: March 1, 2024).
28. West, D.M. *et al*. (2024) The politics of AI: Chatgpt and political bias, Brookings. Available at: https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/ (Accessed: March 1, 2024).
29. Elon Musk, Plaintiff, v. Samuel Altman, Gregory Brockman, OpenAI, Inc., OpenAI, LP, OpenAI, LLC, OpenAI GP, LLC, OpenAI Opco, LLC, OpenAI Global, LLC, OAI Corporation, LLC, OpenAI Holdings, LLC, and DOES 1 through 100, Defendants. Case CGC-24-612746, February 29, 2025. https://www.courthousenews.com/wp-content/uploads/2024/02/musk-v-altman-openai-complaint-sf.pdf (Accessed: March 1, 2024)
30. Li, H; Guo, D; Fan, W; Xu, M; Huang, J; Meng, F; Song, Yangqiu. *Multi-step Jailbreaking Privacy Attacks on ChatGPT.* arXiv:2304.05197v2, pp. 1-9: 2023.
31. Our approach to alignment research, 2023. OpenAI. https://openai.com/blog/our-approach-to-alignment-research (Accessed: August 9, 2023).
32. Hendrycks, D. *et al*. (2022) *What would Jiminy Cricket do? towards agents that behave morally, arXiv.org*. Available at: https://arxiv.org/abs/2110.13136 (Accessed: March 1, 2024).
33. Zou, A. *et al*. (2023) *Papers with code - representation engineering: A top-down approach to AI transparency, Representation Engineering: A Top-Down Approach to AI Transparency | Papers With Code*. Available at: https://paperswithcode.com/paper/representation-engineering-a-top-down (Accessed: 02 March 2024).

Author

Ronan Cheng-Slater is a Junior at the Hopkins School and has studied computer-aided drug design, biomaterials, and programming. He hopes to combine his passions for computing and the sciences in his future studies.