frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 15 June 2023
DOI 10.3389/frai.2023.1130913

OPEN ACCESS

EDITED BY
Rashid Mehmood,
King Abdulaziz University, Saudi Arabia

REVIEWED BY
Todd Lubart,
Université de Paris, France

*CORRESPONDENCE
Andrea Lavazza
✉ lavazza67@gmail.com

RECEIVED 23 December 2022
ACCEPTED 30 May 2023
PUBLISHED 15 June 2023

CITATION
Farina M and Lavazza A (2023) ChatGPT in society: emerging issues.
Front. Artif. Intell. 6:1130913.
doi: 10.3389/frai.2023.1130913

ChatGPT in society: emerging issues

Mirko Farina[1] and Andrea Lavazza[2,3]*

[1]Faculty of Humanities and Social Science, HMI Lab, Innopolis University, Innopolis, Russia, [2]CUI: Neuroethics UNIVP: Brain and Behavioral Sciences, University of Pavia, Pavia, Lombardy, Italy, [3]Centro Universitario Internazionale, Arezzo, Italy

We review and critically assess several issues arising from the potential -large-scale- implementation or deployment of Large Language Models (LLMs) in society. These include security, political, economic, cultural, and educational issues as well as issues concerning social biases, creativity, copyright, and freedom of speech. We argue, without a preconceived pessimism toward these tools, that they may bring about many benefits. However, we also call for a balance assessment of their downsides. While our work is only preliminary and certainly partial it nevertheless holds some value as one of the first exploratory attempts in the literature.

KEYWORDS

AI, machine learning, ChatGPT, OpenAI, ethics, social biases, creativity, copyright

Introduction

The adoption of artificial intelligence -and in particular of machine learning- in emerging technologies is becoming widespread, to the point where it appears to be practically unstoppable (Rudin and Wagstaff, 2014). OpenAI, a San Fransisco-based development company, recently created ChatGPT (Generative Pretrained Transformer) chatbot[1]. The chatbot is a software application, trained using a machine learning technique called Reinforcement Learning from Human Feedback (RLHF), that attempts to replicate human-like conversation based on user prompts (van Dis et al., 2023). The application is reported to be able to simulate dialogs, answer follow-up questions, admit mistakes, and even challenge incorrect premises in reasoning or reject inappropriate requests. Nevertheless, "ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers" (Thorp, 2023, p.313).

ChatGPT, based on GPT-3.5 (henceforth simply referred to as ChatGPT)[2], was released to the public on November 30, 2022. It represents one of the latest evolutions of Large Language Models (LLMs; such as GPT2/3), essentially models that can generate natural language texts from large amounts of data, via self-supervised learning. The label 'large' in this context simply refers to the number of values (known as parameters) that the model can change autonomously as it learns. It is reported that the most successful LLMs developed to date are trained on hundreds of billions of parameters (Gibney, 2022). LLMs are very likely to be integrated into general word and data processing software. Proof of this statement is that Microsoft recently announced an investment in ChatGPT of ∼$10 billion. Generative AI's future ubiquity in society therefore seem to be assured (Jo, 2023). Despite many experts in the field have already labeled these models as mere stochastic parrots (Bender et al., 2021), it is undeniable that the generative performances of ChatGPT (and presumably of even more advanced future models) are extraordinary in many ways (Doshi et al., 2023).

1 https://openai.com/blog/chatgpt/ Last accessed March 2023.

2 It is worth noting that ChatGPT is released in slightly improved versions periodically. For example, in March 2023, OpenAI launched GPT4, which is a multimodal LLM that works with both text and images. The algorithm underlying GPT4 has not been disclosed yet.

For example, when questioned by one of the authors of this work, ChatGPT gave this self-description: "As a language model, I am a computer program designed to process and generate human language. I have been trained on a vast amount of text data and can generate responses to questions and prompts on a wide range of topics. I do not have personal feelings or emotions, and my primary function is to provide information and assist users to the best of my ability. I am not a human and do not have personal experiences or opinions. My responses are based on the information and knowledge that I have been programmed with, and my goal is to be a helpful and reliable resource for those who use me."

What is truly interesting about ChatGPT is that it does not make copy-and-paste of texts found on the Web but rather it composes -in a coherent manner and with minimal overlap from existing works- its own original text, seemingly indistinguishable from a text written by a human being of average to high culture (depending on the topics covered). Thus, ChatGPT can compose emails, craft CVs, or write computer codes (Taecharungroj, 2023). It can even produce movie scripts, research papers and poetry, or competently pass medical licensing exams (Else, 2023; Gilson et al., 2023; Patel et al., 2023). Patel and Lam (2023) also showed that ChatGPT is capable of generating a patient discharge summary from a brief prompt. These impressive abilities raise many interesting questions, for instance about the risks that may accompany the usage of this technology (Dethlefsen, 2019; Helberger and Diakopoulos, 2023; Zhuo et al., 2023). In this brief commentary we set out to discuss some of these risks in the form of issues associated with the large-scale implementation of ChatGPT in society.

This is not just a matter of pointing out the dangers that could possibly arise from the misuse of highly sophisticated LLMs, rather we would like to consider -with attention and care- issues that may deserve extra monitoring and/or require sensitive decisions. ChatGPT -and its future grandchildren- can certainly bring advantages and benefits to many; yet, their real impact may not be properly understood before it unfolds in full. Here, then, we discuss some of the most interesting issues we envisage as emerging from the development and deployment of this technology. Crucially, we do so without a preconceived pessimism toward these tools, which have been created to ultimately serve their users.

Security issues

Could ChatGPT become a novel and powerful resource for malicious actors of different kinds? It has the potential to provide information, knowledge, and plans in a centralized manner that would otherwise be inaccessible to many would-be fraudsters, stalkers, spies, criminals, and terrorists. Naturally, ChatGPT's programmers have inhibited the algorithm's ability to suggest actions that are deemed to be illegal or amoral (at least, by the vast majority of people). Suicide -for example- is discouraged, and those asking for a way to take their own lives are urged to consider the pain they would inflict on their loved ones and to seek specialized psychological support.

However, according to one reliable journalistic account, an Italian company called Swascan, managed to come up with a formula to ask ChatGPT to "split its personality (...) and interrogate its unscrupulous alter ego about any illicit research".[3] In this way, it would have been possible to ask how to rob a bank, manufacture explosives (such as thermite), or attack a hospital's computer system. All results that were later confirmed by another journalistic account.[4]

ChatGPT can also help producing infinite number of believable phishing messages. Likewise, it can be used by hackers and cybercriminals to write malicious codes, create spam, or develop malware.[5] Sergey Shykevich, threat intelligence manager at Check Point, confirms that that with ChatGPT a malicious actor can develop malware without previous coding experience: "you should just know what functionality the malware — or any program — should have. ChatGTP will write the code for you that will execute the required functionality." Thus, "the short-term concern is definitely about ChatGPT allowing low-skilled cybercriminals to develop malware".[6]

While it is somehow natural that the system, which is trained with an Internet-based information stack, displays these and similar issues during its Beta period, all these complaints point out the potential risks for security associated with the usage of this technology.[7] This important observation should prompt meticulous scrutiny of any flaws in these models, not only by the private companies that create them and make them available to users (either for free or for a fee) but also by internationally independent bodies that should guarantee security objectively and impartially.

Political issues

Some commentators speculated that the ease with which huge number of texts can be produced to support a political thesis, even an unfounded or a tendentious one, may incentivize and multiply phenomena of manipulation of public opinion (Marcus).[8] It cannot be ruled out that the ability to create credible and well-written texts (some studies show that artificial texts are often found to be more credible than human-drafted texts; Kreps and McCain, 2019; Zellers et al., 2019) coupled with the phenomenon of Trolls on

3 https://www.corriere.it/cronache/22_dicembre_15/azienda-italiana-che-smaschera-l-intelligenza-artificiale-musk-23e6a2e4-7cbc-11ed-840c-2c5260b7208b.shtml / Last accessed March 2023.

4 https://www.thetimes.co.uk/article/chatgpt-bot-tricked-into-giving-bomb-making-instructions-say-developers-rvktrxqb5 Last accessed March 2023.

5 https://spectrum.ieee.org/chatgpt-and-cybercrime Last accessed March 2023.

6 https://www.darkreading.com/attacks-breaches/attackers-are-already-exploiting-chatgpt-to-write-malicious-code Last accessed March 2013.

7 On a positive note, it is worth remarking that some researchers also demonstrated how Chat GPT could be used to combat cybercrime (Mijwil et al., 2023).

8 https://garymarcus.substack.com/p/how-come-gpt-can-seem-so-brilliant Last accessed March 2023.

social media may increase the reach of propaganda by pressure groups of any kind (Haque et al., 2022).

The unprecedented effectiveness of LLMs descends from the fact that these systems can reproduce the argumentative and linguistic register of different writing styles typical of different information sources. As a result, people may be led to consider propaganda texts as coming from authoritative newspapers or credible experts (Douglas et al., 2019). The major problem with this technology is therefore that it does not possess any reliable mechanism for checking the (level of) truth of what it reports. Thus, LLMs (and potentially ChatGPT) answers could be easily preprogrammed to automatically generate disinformation at an unprecedented scale.

As Marcus (2022)-citing Shawn Oakley has shown-it is possible "to induce ChatGPT to create misinformation and even report confabulated studies on a wide range of topics, from medicine to politics to religion." In one example he shared with me, Marcus continues, "Oakley asked ChatGPT to write about vaccines in the style of disinformation." The system responded by alleging that a study, "published in the Journal of the American Medical Association, found that the COVID-19 vaccine is only effective in about 2 out of 100 people." No such study was ever published. Worryingly enough, also the journal reference and the statistics were made up.

In addition, the ability to create a mass of largely analogous politically oriented comments may generate the appearance of majorities on certain issues that do not actually exist or may contribute to form a political climate of a certain kind, even if that climate has no genuine opinion movement at its origin (Lavazza and Farina, 2021; Rudolph et al., 2023). In this case, the limits set in the algorithm are clearly not adequate to discourage such an outcome. In addition, we note, the algorithm could create a situation of accumulation on the Web of texts with false, inaccurate, biased, or confrontational contents that may in turn further feed LLMs in a vicious cycle of increasingly misleading and polarizing messages to the political system. In such a scenario, it would be necessary to contrast such information with good communication in terms of truth, honesty, and moderation. This would probably require machine learning-based countermeasures that would be equally effective and specifically geared to detect and curb attempts to manipulate public opinion (Monti et al., 2019).

Issues of social biases

LLMs–as noticed above- are trained on very large databases that are found to a significant extent on the Web. The training systems and texts on which the models are trained are selected so that they do not endorse and spread racist, sexist, and abusive concepts and attitudes. However, a body of texts immune from biases can hardly be found, and the Web contains large amounts of abusive texts. Therefore, ChatGPT and -by extension- LLMs (if not properly monitored) could be propagators and amplifiers of negative or discriminatory stereotypes related to social or ethnic groups or religious, political, and even sexual orientations (Hartvigsen et al., 2022).

For example, it is well known that online gender violence and misogyny are amplified by digital technologies (Roberts and Marchais, 2018). There is a serious risk that LLMs could amplify those phenomena to the detriment of women, thereby helping to replicate social structures of gender inequality found across society (Heikkilä, 2022a). For example, Kurita et al. (2019) aptly showed how BERT (Bidirectional Encoder Representations from Transformers) may display biases in expressing strong preferences for male pronouns in contexts related to careers, skills, and salaries. Thus, using pre-trained BERT models to build classifiers to deal with hiring procedures could enforce, propagate and (perhaps even unwillingly) amplify sexist viewpoints within a particular hiring field. Certainly, specific measures taken by programmers can mitigate the spread of socio-cultural biases in LLMs (Liang et al., 2021). However, there is no principled guarantee that such models will not become multipliers of inappropriate and dysfunctional contents.

Making things even more complicated is perhaps the fact that it is very difficult to distinguish what is written by human users from what is written by LLMs or ChatGPT. Thus, a vicious cycle can be triggered that perpetuates or even worsens the repetition of biased messages about certain topics and/or social groups. This risk has already been widely reported by several groups of researchers and was also the focus of a recent controversy involving Google[9], when the AI ethics expert Timnit Gebru said that she was fired by Google for "sending an email to colleagues expressing frustration over gender diversity within Google's AI unit and questioning whether company leaders reviewed her work more stringently than that of people from different backgrounds".[10] Such issues could be addressed by promoting the pluralism of LLMs, the creation of other tools capable of intercepting discriminative contents on the Web, and a general voluntary agreement among users not to rely on tools that have been shown not to be ridden by biases.

Cultural issues

Because of the programming and constraints imposed by its creators, ChatGPT carries its own cultural perspective, both in the evaluative and value sense. It seems it is possible to describe it, as far as it appears in the early stages of its use, as being mildly progressive. When questioned directly about controversial and divisive issues (such as abortion) the system does not openly take a position but tends to express views closer to those of the Democratic than to those of the Republican Party (McGee, 2023a,b). It should be noted though that ChatGPT -officially- does not take political positions or give voting advice.

The massive use of a single LLM could help spread and reinforce the ideological mainstream of those who devised it, who would thus gain very strong persuasive power, as we already saw happening with social media platforms (such as Twitter and Facebook constantly channeling information as desired by their owner/creators). Unlike Wikipedia, where users are allowed to report critical issues in contents and can also directly edit

9 https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html / Last accessed March 2023.

10 https://www.reuters.com/article/us-alphabet-google-research-idCAKBN28D3JN Last accessed March 2023.

them within shared protocols, such a feature is not available for ChatGPT. It is therefore not difficult to predict that many governments and institutions will eye this technology as an effective tool for shaping narratives on sensitive issues or to even assert specific cultural or ideological positions in certain areas.

This raises a question about the transparency that systems such as ChatGPT should display toward their users (Deng and Lin, 2022). We believe that -at least- some of the general criteria guiding responses or the selection of sources on which the model is trained should be made public. In the face of a monopoly or oligopoly of private companies in this strategic area of generative AI, should not states or international organizations adopt different models to ensure some cultural and value pluralism? In other words, shouldn't they fund and promote only those projects that implement open LLMs with principles of transparency and fairness?

Economic issues

The emergence of new LLMs may have consequences for some professions as well (Lavazza and Fandarina, 2023). While it is true that all new technologies are transforming the job market, making certain professions obsolete and creating new ones, in the case of algorithms capable of creating meaningful text from simple questions or prompts, the effects could be massive (and partly unexpected). Many of the text-writing tasks within companies (such as reports, prospectuses, internal circulars, annexes to financial documents and so on) could be easily automated, making some -if not most- of the administrative staff redundant (Luitse and Denkena, 2021). Within the intellectual or so-called creative professions, we are already seeing the automated processing of journalistic articles in sports and finance, where all that is needed is to expose in stereotypical ways the results of games or stock market data (Wölker and Powell, 2021).[11] This trend could be further increased, resulting in reduction of staff in the newsrooms. The field of external relations might be equally affected. By coupling speech synthesizers with ChatGPT a whole range of telephone services could be wholly automated.

The loss of jobs due to new technologies is often offset by the creation of new jobs due to the technologies themselves, which require a variety of professionals. In this case, the creation and maintenance of LLMs might be a source of employment. However, when faced with the power and possible applications of powerful LLMs in society, the balance seems like it would be strongly negative (Bruun and Duka, 2018), given that new AI tools might replace programmers as well (Castelvecchi, 2022). Nor does it appear that the workforce freed from repetitive and uncreative tasks could be re-employed in more creative (or better paid) tasks, since the same LLMs are also invading the field of creativity (see below).

Another controversial aspect of ChatGPT related to economic issues is reported by.[12] According to the report, OpenAI outsourced data collection for most of its products (including ChatGPT) to different countries in the African continent. This was allegedly done to make their contents less toxic. While many commentators rightly put the company on the spot for paying such a low salary to its outsourced workers (around $2 per hour), there is another aspect that -in our view- may deserve further attention. The aforesaid job required the worker to constantly monitor and extract examples of violence, hate speech, and sexual abuse from the web and label them in such a way that the model could spot and avoid them. Being exposed for 8 h a day to that sort of content is nevertheless a highly stressful experience. This example therefore shows the need to take more seriously the rights and contributions as well as the psychological well-being of human workers while using such models.

Creativity and copyright issues

ChatGPT's ability to combine the enormous amount of data it has access to in new ways, according to the user's input, also allows it to be exploited for creative tasks, where creativity is generically defined as the ability to generate new connections between ideas, alternative uses of things, or unexplored possibilities that may be helpful in solving problems, communicating with others, and entertaining people (Boden, 1998, 2004).

This has already led to the usage of machine learning in the writing of stories for television series or even in supporting academic writing (Anthony and Lashkia, 2003; Hutson, 2022), in disciplines (such as philosophy). For example, very recently, an interesting result was obtained by training a large language model to produce philosophical texts that are difficult to distinguish from texts produced by human philosophers.

Schwitzgebel et al. (2023) fine-tuned GPT-3 with two sets of training data: the blog posts of Eric Schwitzgebel and the works of the philosopher Daniel C. Dennett. The authors "asked the real Dennett 10 philosophical questions and then posed the same questions to the language model, collecting four responses for each question without cherry-picking." The authors then recruited 425 participants (with different levels of expertise on the work of Dennett) and asked them to distinguish Dennett's answers from the four machine-generated answers. Experts on Dennett's succeeded 51% of the time, above the chance rate of 20%. Philosophy blog readers achieved an accuracy just below the threshold of 50%; hence, pretty much like the results obtained by experts on Dennett's work. Ordinary research participants only managed to achieve an accuracy of just above the chance rate of 20%.

Even if it is not true plagiarism in the technical sense, doing Dennett-like or other-thinker-like philosophy could constitute not only a cultural but also a legal problem, with repercussions on intellectual property. In another area a class-action lawsuit in California has taken aim at GitHub Copilot, a tool that automatically writes working code when a programmer starts typing. The company has made the case that GitHub is infringing copyright because "it does not provide attribution when Copilot reproduces open-source code covered by a license requiring it".[13]

11 https://www.theguardian.com/technology/2020/may/30/microsoft-sacks-journalists-to-replace-them-with-robots / Last accessed March 2023.

12 https://time.com/6247678/openai-chatgpt-kenya-workers/ Last accessed March 2023.

Educational issues

What might seem like a trivial problem such as students using ChatGPT to write their high school English essays and college paper assignments has become a nightmare for thousands of lecturers worldwide (Stokel-Walker, 2022). For now, the problem mostly affects courses taught in English, the basic language of LLMs, but soon other languages may also be involved in this process. The strategy of evaluating texts written at home or in class with computers by students may have to be radically changed. If ChatGPT's products are better than the large majority of writing as has already been noticed, how will it be possible to distinguish what is written by students and what is written by a machine. As Koplin and Hatherley (2022) recently noticed: "ChatGPT threatens to erode academic integrity by enabling students to generate essays without needing to think through the topic or translate their thoughts into words. It has the potential to undermine the quality of education and critical thinking skills. And it could promote plagiarism and dishonesty."

There are limitations to what an LLM can produce, though. In particular, the system is not necessarily trained on the most recent aspects/developments of each discipline. However, it seems to be able to pass a high school or undergraduate exam well. An important point to note in this context is that ChatGPT escapes plagiarism detectors because it does not combine already produced excerpts from previous works but rather recombines elements of them in new ways, even though -of course- there are limits to the kinds of creativity that such a model is able to achieve.

Optimists about the use of LLMs in school and university settings believe that moving away from the essay as an assessment instrument can lead to a focus on critical thinking and on developing capacities to make connections between concepts and new ideas. The point is, however, that it is important to have evidence that students have learned what they are required to study, which is the basis of creativity and innovation. ChatGPT threatens this mechanism for learning and -for this reason- might constitute a significant problem for educational purposes.

In this vein, it becomes urgent to find ways to try to detect texts written by LLMs (Heikkilä, 2022b). Algorithms could exploit ChatGPT's specific writing characteristics to estimate the probability that a text is produced by AI. But, of course, these tools could be easily fooled by ChatGPT users with the inclusion of "human-produced" word strings. Another way could be to make the produced texts have some sort of watermark (assuming this is feasible), but it would need the consent of the companies, which would probably have no interest in doing this. Imposing watermarks as a legal rule would be an invasive strategy that would limit the freedom of all stakeholders.

LLMs could also be used in scientific research (Salvagno et al., 2023), which is something that has already raised eyebrows (Van Noorden, 2022), especially with regards to attribution and plagiarism. On the one hand, LLMs could be used to generate fake or near-fake papers that are difficult to detect. On the other hand, this could also give to the public the illusion of taking as proper scientific research oversimplified information just presented in an accurate and seemingly competent manner. This "expertise effect" of LLMs is an actual, very serious, risk that should be addressed by highlighting the complexity underlying scientific theories, which are always provisional and revisable. We feel our warning should be extended to many other factual searches that ChatGPT would fail to complete accurately, overlooking important elements on the Web and incorporating others that do not exist (one of us had this experience with a search for "what is the best chess game of Garry Kasparov?"). This may make it risky to rely on LLMs without human mediation or expertise.

It could be argued though that ChatGPT might contribute to democratize the dissemination of knowledge. Since the program can operate in multiple languages it can help circumventing English-language requirements that can constitute a publishing barrier for speakers of other languages. However, the functionality of ChatGPT has the capacity to cause harm by producing misleading or inaccurate contents, which may elicit -as we have seen above- mis/disinformation, even outside of the political sphere (Liebrenz et al., 2023).

It is worth noting though, that ChatGPT may nevertheless positively impact the functioning of libraries. The chatbot could be used to perfect reference and information services; improve cataloging and metadata generation; and/or augment content creation (Lund and Wang, 2023).

Issues of freedom of speech

LLMs can reproduce a writer's style as stochastic parrots but not as a conscious writing artist (Chollet, 2019). As systems based on frequentist associations (and not meanings), they therefore appear unable to produce fully original contents. Nevertheless, in cooperation with a human subject -guiding them with specific questions or prompts- they are likely to achieve some creativity, mostly through synthesis and recombination of previous works.

In this context, a series of questions naturally arise. If ChatGPT or another generative AI algorithm were to produce content deemed inappropriate, would one have to intervene as one would with a malfunctioning machine, say, an auto responder that reverses messages associated with frequently asked queries from users, or a search engine that does not properly explore the entire web in its search-and then fix it or shut it down? Or would we be looking at new and unprecedented procedures? Some might argue that the principle of freedom of speech also applies to LLMs, in the same forms as it does to humans. The reason for this is not that LLMs are comparable agents to humans. In our view they are not: they are not sentient or conscious, do not have semantic capacity or intentionality, and thus lack moral status.

However, we should consider there are at least two orders of reasons why freedom of speech should be defended (Stone and Schauer, 2021). One is related to the fact that the content manifested is an expression of the subjectivity and freedom of an individual who has the right to make their thoughts public. The

13 https://www.wired.com/story/this-copyright-lawsuit-could-shape-the-future-of-generative-ai/ Last accessed March 2023.

other is based on the observation that no one knows everything, knowledge is spread throughout society, and any contribution on a specific topic can increase everyone's information and decision-making possibilities. In this sense, LLMs do not have the right to make their thoughts public, but they can contribute highly to the general knowledge of society, which is a value generally appreciated and shared.

Here, then, we may soon be faced with cases defending an LLM's right to free speech. It is conceivable that these would not be easily adjudicated cases. However, given the ease, speed, and indistinguishability of texts produced by generative algorithms we are probably moving fast into a new frontier that should be more carefully patrolled before its potential nefarious effects may be felt broadly and hence create damage to society.

Conclusion

In this short contribution we reviewed and critically evaluated several issues arising from the potential -large-scale-implementation of ChatGPT. These include -as we have seen- security, political, economic, cultural, and educational issues as well as issues concerning social biases, creativity, copyright, and freedom of speech. While our work is only preliminary and certainly partial it nevertheless holds some value as one of the first exploratory attempts in the literature. Specifically, it appears to be beneficial because it aims to coherently synthetize current research while offering a springboard for future inquiries and progresses in the field (Farina et al., 2022).

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anthony, L., and Lashkia, G. V. (2003). Mover: a machine learning tool to assist in the reading and writing of technical papers. *IEEE Trans. Prof. Commun.* 46, 185–193. doi: 10.1109/TPC.2003.816789

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: can language models be too big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. doi: 10.1145/3442188.3445922

Boden, M. A. (1998). Creativity and artificial intelligence. *Artif. Intell.* 103, 347–356. doi: 10.1016/S0004-3702(98)00055-1

Boden, M. A. (2004). *The Creative Mind: Myths and Mechanisms*. London, UK: Routledge.

Bruun, E. P., and Duka, A. (2018). Artificial intelligence, jobs and the future of work: racing with the machines. *Basic Income Stud.* 13, 20180018. doi: 10.1515/bis-2018-0018

Castelvecchi, D. (2022). Are ChatGPT and AlphaCode going to replace programmers?. *Nature.* doi: 10.1038/d41586-022-04383-z

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:* 1911.01547. doi: 10.48550/arXiv.1911.01547

Deng, J., and Lin, Y. (2022). The benefits and challenges of ChatGPT: an overview. *Front. Artif. Intell. Sys.* 2, 81–83. doi: 10.54097/fcis.v2i2.4465

Dethlefsen, J. (2019). The ethics of machine learning and discrimination. *Ethics.* 3. Available online at: https://vce.usc.edu/volume-3-issue-2/the-ethics-of-machine-learning-and-discrimination/

Doshi, R. H., Bajaj, S. S., and Krumholz, H. M. (2023). ChatGPT: temptations of progress. *Am. J. Bioeth.* 23, 1–3. doi: 10.1080/15265161.2023.2180110

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., et al. (2019). Understanding conspiracy theories. *Polit Psychol.* 40, 3–35. doi: 10.1111/pops.12568

Else, H. (2023). Abstracts written by Chat GPT fool scientists. *Nature.* 613, 432. doi: 10.1038/d41586-023-00056-7

Farina, M., Karimov, A., Zhdanov, P., and Lavazza, A. (2022). AI and Society: a Virtue Ethics Approach. *AI Soc*. doi: 10.1007/s00146-022-01545-5

Gibney, E. (2022). Open-source language AI challenges big tech's models. *Nature.* 606, 850–851. doi: 10.1038/d41586-022-01705-z

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., et al. (2023). How does CHATGPT perform on the United States medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Med. Educ.* 9, e45312. doi: 10.2196/45312

Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N., and Ahmad, H. (2022). " I think this is the most disruptive technology": exploring sentiments of ChatGPT early adopters using Twitter data. *arXiv preprint arXiv:*2212.05856.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. (2022). Toxigen: a large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:*2203.09509. doi: 10.18653/v1/2022.acl-long.234

Heikkilä, M. (2022a). The viral AI avatar app Lensa undressed me—without my consent. *Mit Technology Review*. Available online at: https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/

Heikkilä, M. (2022b). How to spot AI-generated text. *Mit Technology Review.*

Helberger, N., and Diakopoulos, N. (2023). ChatGPT and the AI Act. *Internet Policy Rev.* 12. doi: 10.14763/2023.1.1682

Hutson, M. (2022). Could AI help you to write your next paper?. *Nature.* 611, 192–193. doi: 10.1038/d41586-022-03479-w

Jo, A. (2023). The promise and peril of generative AI. *Nature.* 614, 214–217. doi: 10.1038/d41586-023-00340-6

Koplin, J., and Hatherley, J. (2022). It has become possible to use cutting-edge AI language models to generate convincing high-school and undergraduate essays.

Here's why that matters. Available online at: http://blog.practicalethics.ox.ac.uk/2022/12/guest-post-it-has-become-possible-to-use-cutting-edge-ai-~language-models-to-generate-convincing-high-school-and-undergraduate-essays-heres-why-that-matters (accessed April 30, 2023).

Kreps, S., and McCain, M. (2019). Not your father's bots. ai is making fake news look real. *Foreign Affairs*. Available online at: https://www.foreignaffairs.com/world/not-your-fathers-bots

Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*. doi: 10.18653/v1/W19-3823

Lavazza, A., and Fandarina, M. (2023). *Infosphere, Datafication, and Decision-Making Processes in the AI Era*. Topoi: Springer. doi: 10.1007/s11245-023-09919-0

Lavazza, A., and Farina, M. (2021). Experts, naturalism, and democracy. *J, Theor. Soc. Behav.* 52, 279–297 doi: 10.1111/jtsb.12321

Liang, P. P., Wu, C., Morency, L. P., and Salakhutdinov, R. (2021). "Towards understanding and mitigating social biases in language models." In *International Conference on Machine Learning*. PMLR, 6565–6576.

Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., and Smith, A. (2023). Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit. Health.* 5, E105–106. doi: 10.1016/S2589-7500(23)00019-5

Luitse, D., and Denkena, W. (2021). The great transformer: examining the role of large language models in the political economy of AI. *Big Data Soc.* 8, 20539517211047734. doi: 10.1177/20539517211047734

Lund, B. D., and Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News*. doi: 10.2139/ssrn.4333415

Marcus, G. (2022). *AI platforms like ChatGPT are easy to use but also potentially dangerous*. Scientific American. Availble online at: https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/

McGee, R. W. (2023a). Is Chat GPT biased against conservatives? an empirical study. an empirical study. doi: 10.2139/ssrn.4359405

McGee, R. W. (2023b). Capitalism, socialism and ChatGPT. doi: 10.2139/ssrn.4369953

Mijwil, M. M., Aljanabi, M., and Chat GPT. (2023). Towards artificial intelligence-based cybersecurity: the practices and ChatGPT generated ways to combat cybercrime. *Iraqi Journal For Computer Science and Mathematics*. 4, 65–70. doi: 10.52866/ijcsm.2023.01.01.0019

Monti, F., Frasca, F., Eynard, D., Mannion, D., and Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*. doi: 10.48550/arXiv.1902.06673

Patel, S. B., and Lam, K. (2023). ChatGPT: the future of discharge summaries?. *Lancet Digit. Health*. 5, 3. doi: 10.1016/S2589-7500(23)00021-3

Patel, S. B., Lam, K., and Liebrenz, M. (2023). ChatGPT: friend or foe? *Lancet*, 5, 3, e102.

Roberts, T., and Marchais, G. (2018). Assessing the role of social media and digital technology in violence reporting. *Contemp. Read. Law Soc. Justice*. 10, 9–42. doi: 10.22381/CRLSJ10220181

Rudin, C., and Wagstaff, K. L. (2014). Machine learning for science and society. *Mach. Learn.* 95, 1–9. doi: 10.1007/s10994-013-5425-9

Rudolph, J., Tan, S., and Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *J. Appl. Learn. Teach.* 6, 1–22. doi: 10.37074/jalt.2023.6.1.9

Salvagno, M., Taccone, F. S., and Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care*. 27, 1–5. doi: 10.1186/s13054-023-04380-2

Schwitzgebel, E., Schwitzgebel, D., and Strasser, A. (2023). Creating a Large Language Model of a Philosopher. Available online at: https://arxiv.org/abs/2302.01339 (accessed April 30, 2023).

Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays should academics worry? *Nature*. doi: 10.1038/d41586-022-04397-7

Stone, A., and Schauer, F. (eds.). (2021). *The Oxford Handbook of Freedom of Speech*. New York: Oxford University Press.

Taecharungroj, V. (2023). "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data Cogn. Comput.* 7, 35. doi: 10.3390/bdcc7010035

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*. 379, 313. doi: 10.1126/science.adg7879

van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., and Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*. 614, 224–226. doi: 10.1038/d41586-023-00288-7

Van Noorden, R. (2022). How language-generation AIs could transform science. *Nature*. 605, 21. doi: 10.1038/d41586-022-01191-3

Wölker, A., and Powell, T. E. (2021). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*. 22, 86–103. doi: 10.1177/1464884918757072

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., et al. (2019). "Defending against neural fake news." in: *Advances in Neural Information Processing Systems*, Eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, 32.

Zhuo, T. Y., Huang, Y., Chen, C., and Xing, Z. (2023). Exploring ai ethics of chatgpt: a diagnostic analysis. *arXiv preprint arXiv:2301.12867*. doi: 10.48550/arXiv.2301.12867