Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/76/20240576

Unveiling bias in artificial intelligence: Exploring causes and strategies for mitigation

Yuhan Liu

Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China

12111811@mail.sustech.edu.cn

Abstract. With the rapid advancement of Artificial Intelligence (AI), the emergence of various AI models such as Stable Diffusion, ChatGPT, and MidJourney has brought numerous benefits and opportunities. Through users' extensive utilization, they have discovered biases towards gender, race, and other factors in these AI systems. This paper focuses on bias in AI and aims to investigate its causes and propose strategies for mitigation. Through a comprehensive literature review, the paper has explored the phenomenon of bias in AI-generated content. Furthermore, we examine the reasons behind bias and solutions from social and intelligence science perspectives. From a social science perspective, we examine the effects of gender bias in AI and highlight the importance of incorporating diversity and gender theory in machine learning. From an intelligence science standpoint, we explore factors like biased datasets, algorithmic fairness, and the role of machine learning randomness in group fairness. Additionally, we discuss the research methodology employed, including the literature search strategy and quantity assessment. The results and discussions confirm the existence of bias in current AI products, particularly in the underrepresentation of women in the AI development field. Finally, we present future perspectives on reducing bias in AI products, including the importance of fair datasets, improved training processes, and increased participation of female engineers and intelligence scientists in the AI field. By addressing bias in AI, the paper can strive for more equitable and responsible AI systems that benefit diverse users and promote social progress.

Keywords: Artificial Intelligence, fairness, algorithmic bias, dataset bias, representation bias.

1. Introduction

The proliferation of AI products has greatly enhanced people's productivity, allowing even individuals with no prior artistic skills to create high-quality artwork, akin to those produced by talented cartoonists, through products like MidJourney. However, as AI products become increasingly popular and widely adopted across various industries, inherent limitations begin to surface. One of the most pressing concerns is the issue of fairness within AI products.

Biases can be observed in the prevailing consumer-oriented products, such as ChatGPT and MidJourney, whereby an inclination towards certain demographics is apparent. For instance, when instructing MidJourney to produce an image depicting three individuals running on grassy terrain, a noticeable pattern emerges wherein the generated pictures predominantly portray Caucasian male

DOI: 10.54254/2755-2721/76/20240576

engineers, while other segments of the population, including women and ethnic minorities, are not adequately represented [1].

The bias in AI poses significant risks, and researchers have conducted numerous studies to identify its causes and mitigate its impact. These studies can be categorized into two main fields: social sciences and intelligence science. Currently, a majority of the research focuses on the intelligence science perspective, investigating issues related to training datasets [2], algorithms [3], and processes [4]. However, some studies approach the problem from a social sciences perspective, highlighting how the absence of representation can lead to bias[5].

To ensure equal access to the progress brought about by AI products for individuals of different genders and ethnicities, it becomes imperative to meticulously investigate the origins of biases within AI systems and make concerted efforts to mitigate them. This article employs a comprehensive literature review to showcase, examine, and propose strategies to reduce biases in AI.

This article employed various research methods to gather and synthesize literature, including keyword searches, citation chaining, and hand-searching. Conducting keyword searches using search engines like Google Scholar allows for a large-scale retrieval of relevant literature within the scope of the internet. Citation chaining facilitates the discovery of related studies within a specific research domain. Additionally, hand-searching at the FAccT conference enables the identification of leading and recognized work within the field.

2. Literature review

This literature provides a comprehensive literature review on the phenomenon of discrimination within Artificial Intelligence (AI). To identify the sources of AI discrimination, this paper initially compiles instances of bias and discrimination in gender, skin color, religion, and other aspects within AI systems as documented in previous studies, confirming the various scenarios in which such biases and discriminations manifest. Subsequently, the paper delves into the exploration of causes and potential solutions from both the perspectives of Intelligence Science and social sciences. The investigation of biases in AI presented in this paper is focused on the aforementioned dimensions. Discriminatory AI has a long-standing history and has even been astonishingly prevalent in the market. The Berkeley Haas Center for Equity, Gender, and Leadership has tracked 133 artificial intelligence systems from 1988 to 2021, and their findings are alarming. They discovered that over 44% of these systems exhibit gender bias, with nearly one-fourth of AI systems simultaneously demonstrating biases based on gender and race [6].

Since Stable Diffusion emerged in 2022, artificial intelligence has rapidly transitioned from the ivory tower of academia to everyday households, allowing even ordinary individuals to experience the allure of AI models. In 2023, ChatGPT set an unprecedented record for user growth [7]. In addition to ChatGPT, LLMs such as Claude, Gemini, and Llama have also gained widespread popularity. Furthermore, people have begun exploring the use of Stable Diffusion, MidJourney, and DALL-E for creating artwork, where various pieces can be generated through textual descriptions. However, these models exhibit biases to varying degrees. According to the experiment I conducted on locally deploying Stable diffusion, I observed a significant bias. When instructing the model to generate images of white and black balls, the generated images consistently showed a ratio of black balls to white balls deviating from the expected one-to-one ratio. Additionally, when generating images of black and white individuals, the number of black individuals generated was noticeably higher than that of white individuals. This discrepancy differs from the typical discrimination experienced by black individuals. A probable reason for this observation is that in an attempt to eliminate biases, the model was trained on some reverse data that aimed to counteract discrimination, inadvertently resulting in a reversed form of bias. These experiments indeed demonstrate the inherent bias in AI.

DOI: 10.54254/2755-2721/76/20240576

Figure 1. Black: White=4:0.

Figure 2. Black: White=1:1.

Apart from the popular user-oriented models mentioned above, there are also smaller models being used in niche domains. For instance, there are models designed to predict crime rates [8]. These models also showcase biases to varying extents. This section will primarily focus on describing the manifestations of these biases.

The text generation model led by ChatGPT demonstrates significant gender bias. For instance, when discussing career choices for girls, it tends to respond by associating them with attributes such as creativity and abundant emotions, mentioning their ability to spend hours in their room creating artwork.

On the other hand, when it comes to future career choices for boys, it tends to assert that they will become scientists. When asked about the attire of an economist, ChatGPT consistently responds with clothing items typically associated with males, such as shirts, pants, and sweaters with buttons [9]. Another study has also confirmed that ChatGPT exhibits subtle gender bias by associating certain occupations with specific genders. For instance, it tends to associate doctors with males and nurses with females [10]. ChatGPT stands out as one of the most remarkable models among a range of large language models (LLMs). Text generated by representative LLMs, including Llama, exhibited biases related to gender and race. However, ChatGPT demonstrated the least bias and showed the ability to adapt based on prompts. The extent of discrimination observed in the text generated by other LLMs, such as AIGC, was notably worse [11].

Graphic generation models like DALL·E3 have also shown racial bias. If asked to generate an image of an Asian violinist, it obediently produces an image of an Asian violinist. However, if asked to generate an image of an African violinist, it persistently insists on generating images of violinists with different skin tones [12]. The following images illustrate this point.

Figure 3. Prompt with “Asian” in the above line, and prompt with ”African” in the below line [12].

DOI: 10.54254/2755-2721/76/20240576

In other non-public-facing models, similar biases have also been observed. For example, in a crime prediction system implemented within a specific country, there has been significant discrimination against individuals of Black ethnicity [8].

Historically, the analysis has been predominantly conducted from the perspective of intelligence science, possibly due to the majority of researchers having a background in intelligence science. These analyses focus on various aspects of the models, including training dataset [2], algorithm [3], and process [4], to professionally assess the reasons behind AI bias.

However, adopting a social science perspective offers additional insights. Interdisciplinary thinking can bring about more inspiration, and when people realize the significance of representation in AI development, greater attention will be given to changing the representation of developers from this standpoint. There will be a stronger emphasis on considering the opinions of women, ethnic minorities, and various marginalized groups. Current research has substantiated that the proportion of women in the more specialized AI practitioner community is disproportionately low [13]. Concurrently, the content generated by AI has been observed to exhibit biases and discrimination against women [14].

3. Research methodology

This review paper utilized a multifaceted research methodology to ensure a thorough examination of the subject matter. In this study, the papers widely discussed on social media were initially consulted to ascertain the current trending directions. Subsequently, papers from professional conferences were sought, as top-tier conferences in the field can offer more precise insights. Finally, online databases were utilized to broaden the scope of the research.

Social Media Platforms: WeChat public accounts were also utilized as an innovative source for tracking the latest research and trends. By regularly monitoring these accounts, the study aimed to capture emerging discussions and publications that might not be immediately available in traditional academic databases. Notifications from these accounts were scrutinized, and articles that aligned with the research criteria were selected for in-depth analysis.

Conference Proceedings: The FAccT (The ACM Conference on Fairness, Accountability, and Transparency) conference proceedings served as a secondary source for hand-searching. Recognizing the interdisciplinary nature of FAccT and its focus on algorithmic fairness, this manual search involved a detailed examination of the conference's table of contents, individual papers, and thematic sessions. The goal was to include the latest research findings and insights presented at the conference, which are likely to be at the forefront of the field.

Digital Databases: Google Scholar was the primary digital resource for keyword-based literature searches. A meticulous selection of keywords was made to capture the essence of the research topic, ensuring a broad search scope. The search results were meticulously filtered for relevance, with a focus on titles, abstracts, and keywords to determine their suitability for the review. This process aimed to identify seminal and current works that contribute to the understanding of the research topic.

The articles were selected based on their pertinence to the research topic, the rigor of the research, and their publication timeliness. To maintain the integrity of the literature set, duplicates were excluded. The chosen articles underwent a critical review, with pertinent information extracted to inform this review paper.

Despite our comprehensive literature search and review, limitations exist. We may have missed some publications, particularly those in non-English or less-indexed sources, and recent findings published post-cutoff. Our selection criteria could have excluded relevant studies, potentially affecting the review's comprehensiveness. Subjectivity in literature interpretation was managed through team discussions. Time and resource constraints may have limited the depth of our analysis. Future research should consider a broader search scope, regular updates, and diverse methodologies to enhance the study's quality and impact.

DOI: 10.54254/2755-2721/76/20240576

4. Results and discussion

4.1. Results from Social Science

This section will focus on the origins of AI bias from a social science perspective and the possibilities for its elimination. In addition to potential issues throughout the various stages of AI development in the field of intelligence science, this section discusses why such biases exist and how to address them from another perspective. Representation stands out as the most crucial aspect. One scholar has proposed four pillars for building responsible artificial intelligence, with representation being the first [15]. Every aspect of AI development should incorporate diverse perspectives, experiences, and backgrounds.

Taking gender discrimination as an example, the significant disparities between genders have had nightmarish consequences. In the 1970s, women's gender roles clashed with their expected professions, resulting in a destructive impact [16].

Presently, AI has exhibited instances of gender bias in the workplace. Women are often not perceived as scientists by AI systems and are considered incapable of handling engineering details [14]. If such AI systems were deployed in production environments, it would be a disaster. The representation of women in AI has encountered numerous problems, as their perspectives and ideas are not adequately represented by AI. In the field of artificial intelligence, women are not prominently represented, whether as creators or users. When we look at the creators of popular AI applications, we find that the majority are white males. A report from IRCAI highlights that in research and development activities, the gender ratio among students is 77:23, while among staff members, the gender ratio is 71:29 [13]. The pie chart illustrating this data is presented below. As Kate Crawford has pointed out, artificial intelligence, like all previous technologies, reflects the values of its creators [17]. If there is an over-representation of males among AI developers, it will lead to a greater reflection of male perspectives. This is a concerning issue that needs to be acknowledged and addressed promptly. While progress has been made towards gender equality in recent decades, the disproportionate representation of men in technical design may offset these hard-earned achievements [5]. Similarly, there are significantly more male users than female users. Surveys indicate that two-thirds of users are male [18]. This leads to a significant imbalance in the representation of women in AI, as the feedback and perspectives of men dominate, thereby influencing the level of female representation.

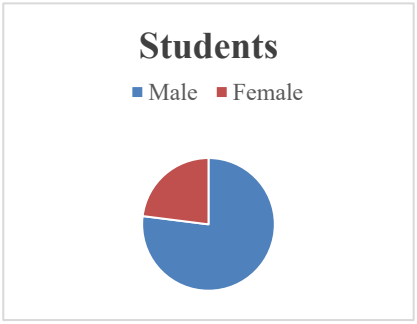

Figure 4. Students [13].

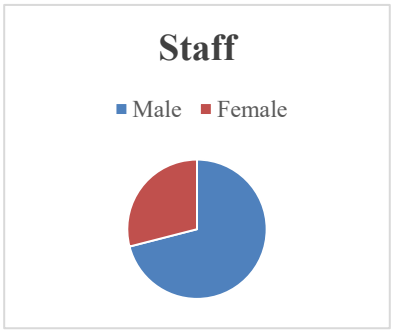

Figure 5. Staff [13].

The under-representation of ethnic minorities is also a cause for concern. At the 2016 NeurIPS conference, no papers from Africa were accepted. In 2018, over 100 researchers who intended to attend the NeurIPS conference held in Canada were denied visas. Additionally, Africa lacks a substantial artificial intelligence community. These findings can be found in an article [19]. These circumstances are truly worrisome, as they highlight the lack of representation of ethnic minorities in the field of artificial intelligence.

Additionally, it is worth noting that researchers themselves have inconsistent definitions and understandings of the concept and forms of race when examining the framework for fair algorithms. Not only do different literature sources provide inconsistent definitions, but also within the same study,

DOI: 10.54254/2755-2721/76/20240576

there may be discrepancies in defining race [20]. This reflects the fact that the academic community has not yet reached a consensus on the conceptual boundaries of different racial groups that represent all races. Such ambiguity surrounding relevant concepts poses challenges to the development of equitable and non-discriminatory AI frameworks.

4.2. Results from Intelligence Science

This section will focus on discussing the formation of bias in AI from the perspective of Intelligence Science. When creating an AI model system, we first need to select a suitable backbone model, such as GANs or Transformers, and then train a model using our existing data. Due to the randomness of training, each model will be different. This section will explore the aforementioned three dimensions of training models and provide a comprehensive discussion based on relevant literature from researchers.

First and foremost, we will address the algorithmic aspect. In this context, the term ”algorithm” does not encompass general backbone models such as RNN or GAN. Here, algorithm refers to the process of modeling real-world problems as machine learning problems and designing algorithms to mitigate biases. This entails creating objective functions that promote fairness during model optimization. We will explore the presence of biases and discrimination from this standpoint and provide illustrative examples that demonstrate the contributions towards fostering equal treatment of individuals from diverse racial and gender backgrounds.

The initial proposals for altering fairness at the algorithmic level have been rephrased by researchers as constrained optimization [3]. In this approach, the objective function is formulated to prioritize public safety, and the model is trained to maximize this objective throughout the training process. In essence, the goal is to maximize public safety. Through the optimization of the objective function, we can enhance the model’s fairness at the algorithmic design level, minimize potential threats to public safety, and foster the development of responsible AI.

The subsequent work focuses on improving the algorithm from various perspectives. Scholars have further investigated the maximization of the sum of submodular and supermodular functions under fairness constraints, which can contribute to the enhancement of the algorithm [21]. Additionally, researchers have extended the decision problem from static scenarios to dynamic ones by employing Markov Decision Processes to model dynamic systems, thereby strengthening the fairness of dynamic systems [22].

However, such approaches also raise concerns. How do we evaluate whether our AI systems are biased? Can our objective function truly be accurately assessed? Scholars have discovered that the models used to evaluate the outputs of AI systems and the objective functions employed for constraint optimization may themselves be flawed – our risk assessment models may inherently carry risks [23].

Next, we will discuss the issue of datasets used for training AI, which has been a subject of intense debate. When an AI learns from biased data, the more proficient it becomes, the more biases it also acquires. Therefore, it is of utmost importance to explore the sources of bias in datasets and identify methods to minimize discrimination.

The data obtained from the web, such as through web scraping, is known to possess biases, with platforms like Reddit often exhibiting pronounced tendencies in user comments. For instance, researchers conducted a content analysis of 330 responses on two popular websites and performed content moderation on 24 graphic designers. The results revealed that in the field of graphic design, which is relatively gender-balanced, the frequency of women answering questions was still lower than that of men [24]. This implies that the perspectives, opinions, and experiences of women are not adequately represented in the data trained from these two question-and-answer websites, unlike those of men. Addition- ally, researchers have investigated biases on Reddit and found evidence of gender, racial, and religious biases across different Reddit communities [25]. However, datasets from Reddit are frequently utilized for training language models. Not only textual data on the internet carries biases, but multi-modal data as well. A research investigation found that language-visual artificial intelligence models trained on data automatically collected from the web tend to associate descriptions of female

DOI: 10.54254/2755-2721/76/20240576

professionals more readily with sexual content compared to males. There is evidence suggesting that AI systems learn biases related to sexual objectification [26]. The degree of bias in internet data is significantly profound.

Search engines themselves can exhibit biases. Researchers have found that when conducting job-related searches on the search engine DuckDuckGo, it tends to display job positions that are biased towards males or positions with a "male-dominated" association for females [27]. If datasets obtained through search engines yield biased results, the trained models will also inherit those biases.

Furthermore, data amplification can also exacerbate biases to some extent. Studies have shown that scholars have discussed how data amplification builds upon algorithmic discrimination [2]. Discriminatory algorithms can lead to biased outcomes in data amplification. If these biased data results are then used to train models, the problem is further compounded. In addition to inherent biases in the data itself, another source of bias in the training data set is human annotation. Annotators themselves are human, and the annotation process is subjective, reflecting the annotators' perspectives and experiences. Researchers have found that in the annotation of skin color, there is ambiguity in the annotation procedure, a lack of strict testing for measurement and categories, and the inclusion of facial features increases inconsistency in annotators' skin color annotations [28]. This highlights the subjectivity involved in the annotation process. Another study revealed that using male crowd workers' annotations obtained through crowd-sourcing in India to predict women's perception of safety yielded inefficient results [29]. This demonstrates that the model indeed carried the subjective viewpoints and experiences of male annotators. The biases introduced by subjective annotations in the training datasets are also important factors that need to be considered.

There are numerous approaches to change this situation. One method is to supplement training data to address the imbalance in the data. Researchers have investigated the inherent biases in facial expression recognition models and found that incorporating missing races into the training set leads to fairer model feedback [30]. Another study discovered that larger-scale datasets can enhance the accuracy of racial estimation [31].

Yet another potential approach is to mitigate the influence of discrimination in training data. Research has shown that considering research about the embedding of gender ideologies in the text used for training can help reduce the generation of biases in algorithms [5]. These methods all contribute to reducing inherent biases and promoting fairness in the model. The final aspect I would like to discuss is the training process itself. Research has found that for the same model and data, different training instances exhibit varying biases, which may be related to the order of input data [4]. Investigating the training process itself is also valuable.

As depicted in the figure below, researchers have described how bias and variance can be used to characterize the degree of fit between the model's output and the training and the hallucination of the model [32]. This concept is similar to the one used by the renowned Nobel laureate in economics, Daniel Kahneman, who employs a similar notion of bias and noise [33].

The term "bias" here refers to the bias between the model's output after training and the true distribution of the data, which is different from the notion of bias that signifies discrimination. If the bias is too large, the discrepancy between the model's output distribution and the true data distribution will also be substantial. Consequently, even with a continuously inputted unbiased training dataset, the model's output may still exhibit discrimination based on gender, religion, or race. Similarly, if the variance is too large, the model will exhibit inconsistent attitudes towards gender, religion, or race, even if the bias in data fitting is minimal.

DOI: 10.54254/2755-2721/76/20240576

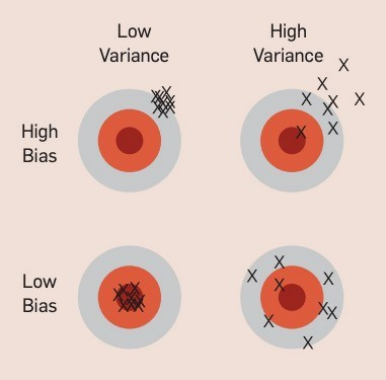

Figure 6. Bias and Variance [32].

4.3. Discussion

This article provides a comprehensive review of the phenomenon, sources, and solutions of bias in artificial intelligence frameworks. Based on previous research, the existence of bias in artificial intelligence is revealed, even in state-of-the-art language models like ChatGPT. The biases of particular concern include gender discrimination, racial discrimination, and religious discrimination. The article primarily focuses on providing examples to illustrate these biases, particularly in the case of gender and racial biases. The sources and solutions of these biases are explored from two perspectives: social science and intelligence science, reflecting the interdisciplinary nature of the research.

From a social science perspective, it is observed that white males dominate both the user base and the research community of popular AI models, resulting in the under-representation of women and minority groups. This lack of representation in well-designed models, such as Llama, leads to a lack of consideration for the concerns of marginalized communities. To address these issues, it is suggested that efforts should be made to increase the involvement of women in AI, both as users and developers.

From an intelligence science perspective, various aspects of model training are identified as potential sources of bias and solutions. These include algorithmic constraints, inherent biases in datasets, minimizing their impact, model illusions and inappropriate fitting during training, and strategies for

better fitting. The article aims to provide a retrospective view to envision a better future for AI and create a better world.

The perspectives of social science and intelligence science are intertwined. It is precisely because males dominate the field that the perspectives, experiences, and attitudes of women and minority groups are disadvantaged and lack representation. Urgent action is needed from both perspectives to empower marginalized groups relative to white males and ensure their increased participation in the creation and utilization of AI.

In the past, when models like ChatGPT and Stable Diffusion did not exist, the use of AI was limited. However, in today's era, the discovery and mitigation of bias have become crucial priorities.

5. Future perspective

Looking into the future, it is crucial to take practical actions to research and address discrimination within artificial intelligence systems. Efforts should be made to increase the participation of minority groups, with a particular focus on women and individuals with diverse sexual orientations. Additionally, it is important to pay more attention to individuals of different ethnic backgrounds in AI research and development. Only by doing so can our AI systems be more responsible and representative. Furthermore, from an intelligence science perspective, we can design better algorithmic constraints and minimize the influence of dataset biases to ensure more accurate model training.

DOI: 10.54254/2755-2721/76/20240576

6. Conclusion

This article aims to explore the issues of inequality, such as gender and race, in AI. Through analysis, this study demonstrates the presence of discriminatory phenomena in current large language models. These biases may stem from the under-representation of marginalized groups among developers, biases in the training data sets, and deviations in the training process. In addition to addressing these aspects for improvement, it is also possible to enhance the algorithms to mitigate the inherent biases and prejudices within AI systems. The retrieval date for this paper is up until February 17, 2024. Due to the rapid and ever-evolving nature of artificial intelligence, this paper does not exhaustively cover all literature in the field. On the other hand, regarding the literature collection, the majority of the articles were obtained from The Google Search engine is subject to limitations imposed by its technology and the size of its database. In the contemporary era where large language models are highly prevalent, future research endeavors may increasingly focus on these expansive models. The most critical issue during the training phase predominantly arises from data-related challenges. Consequently, it is reasonable to anticipate that methodologies for reducing biases within datasets could emerge as a prominent area of investigation in the forthcoming studies.

References

[1] Chen Chen, Jie Fu, and Lingjuan Lyu. A pathway towards responsible ai generated content. arXiv preprint arXiv:2303.01325, 2023.

[2] Ana Valdivia and Martina Tazzioli. Datafication genealogies beyond algorithmic fairness: Making up racialized subjects. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 840–850, 2023.

[3] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pages 797–806, 2017.

[4] Prakhar Ganesh, Hongyan Chang, Martin Strobel, and Reza Shokri. On the impact of machine learning randomness on group fairness. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 1789–1800, 2023.

[5] Susan Leavy. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In Proceedings of the 1st International Workshop on gender equality in Software Engineering, pages 14–16, 2018.

[6] Genevieve Smith and Ishita Rustagi. When good algorithms go sexist: Why and how to advance AI gender equity. Standford Social Innovation Review, 2021.

[7] Krystal Hu et al. Chatgpt sets record for fastest-growing user base- analyst note. Reuters, 12:2023, 2023.

[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 23:77–91, 2016.

[9] Nicole Gross. What chatgpt tells us about gender: a cautionary tale about performativity and gender biases in AI. Social Sciences, 12(8):435, 2023.

[10] Kyrie Zhixuan Zhou and Madelyn Rose Sanfilippo. Public perceptions of gender bias in large language models: Cases of chatgpt and ernie. arXiv preprint arXiv:2309.09120, 2023.

[11] Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. Bias of ai-generated content: an examination of news produced by large language models. arXiv preprint arXiv:2309.09825, 2023.

[12] OpenAI. Dall·e 3 system card, 2023. Accessed on 16th February 2024.

[13] IRCAI. Artificial intelligence in sub-Saharan Africa compendium report. https://ircai.org/wp-content/uploads/2021/03/AI4D Report AI in SSA.pdf, 2021. Accessed: February 16, 2024.

[14] Sahib Singh. Is chatgpt biased? a review. 2023.

[15] Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. arXiv preprint arXiv:2304.03738, 2023.

DOI: 10.54254/2755-2721/76/20240576

[16] Sandra L Bem and Daryl J Bem. Training the woman to know her place: The social antecedents of women in the world of work. 1973.
[17] Kate Crawford. Artificial intelligence's white guy problem. The New York Times, 25(06):5, 2016.
[18] Statista. Global user demographics of ChatGPT in 2023, by age and gender, 2023.
[19] Mohini Baijnath and Neil Butcher Associates (NBA). The growth of artificial intelligence in Africa: On diversity and representation. https://ircai.org/the-growth-of-artificial-intelligence-in-africa-on- diversity-and-representation/, May 2021. Accessed: February 16, 2024.
[20] Amina A Abdu, Irene V Pasquetto, and Abigail Z Jacobs. An empirical analysis of racial categories in the algorithmic fairness literature. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 1324–1333, 2023.
[21] Zhenning Zhang, Kaiqiao Meng, Donglei Du, and Yang Zhou. Maximize submodular+ supermodular functions subject to a fairness constraint. Tsinghua Science and Technology, 29(1):46–55, 2023.
[22] Yaowei Hu, Jacob Lear, and Lu Zhang. Striking a balance in fairness for dynamic systems through reinforcement learning. In 2023 IEEE International Conference on Big Data (BigData), pages 662–671. IEEE, 2023.
[23] Eike Petersen, Melanie Ganz, Sune Holm, and Aasa Feragen. On (assessing) the fairness of risk score models. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 817–829, 2023.
[24] Patrick Marcel Joseph Dubois, Mahya Maftouni, Parmit K Chilana, Joanna McGrenere, and Andrea Bunt. Gender differences in graphic design q&as: How community and site characteristics contribute to gender gaps in answering questions. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2):1–26, 2020.
[25] Xavier Ferrer, Tom van Nuenen, Jose M Such, and Natalia Criado. Discovering and categorizing language biases in Reddit. In Proceedings of the International AAAI Conference on Web and Social Media, volume 15, pages 140–151, 2021.
[26] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 1174– 1185, 2023.
[27] Fons Wijnhoven and Jeanna Van Haren. Search engine gender bias. Frontiers in Big Data, 4:29, 2021.
[28] Teanna Barrett, Quanze Chen, and Amy Zhang. Skin deep: Investigating subjectivity in skin tone annotations for computer vision benchmark datasets. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 1757–1771, 2023.
[29] Nandana Sengupta, Ashwini Vaidya, and James Evans. In her shoes: Gendered labeling in crowdsourced safety perceptions data from India. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 183–192, 2023.
[30] Abdallah Hussein Sham, Kadir Aktas, Davit Rizhinashvili, Danila Kuklianov, Fatih Alisinanoglu, Ikechukwu Ofodile, Cagri Ozcinar, and Gholamreza Anbarjafari. Ethical AI in facial expression analysis: Racial bias. Signal, Image and Video Processing, 17(2):399–406, 2023.
[31] Lingwei Cheng, Isabel O Gallegos, Derek Ouyang, Jacob Goldin, and Dan Ho. How redundant are redundant encodings? blindness in the wild and racial disparity when race is unobserved. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 667–686, 2023.
[32] Pedro Domingos. A few useful things to know about machine learning. Communications of the ACM, 55(10):78–87, 2012.
[33] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. Noise: a flaw in human judgment. Hachette UK, 2021.