12-2023

Perception of Bias in ChatGPT: Analysis of Social Media Data

Abdullah Wahbeh
Slippery Rock University of Pennsylvania

Mohammad A. Al-Ramahi
Texas A&M University-San Antonio, mrahman1@tamusa.edu

Omar El-Gayar
Dakota State University

Ahmed El Noshokaty
California State University - San Bernardino

Tareq Nasralah
Northeastern University

Perception of Bias in ChatGPT: Analysis of Social Media Data

Abdullah Wahbeh
Slippery Rock University
abdullah.wahbeh@sru.edu

Mohammad Al-Ramahi
Texas A&M University-San Antonio
mohammad.abdel@tamusa.edu

Omar El-Gayar
Dakota State University
omar.el-gayar@dsu.edu

Ahmed Elnoshokaty
California State University, San Bernardino
ahmed.elnoshokaty@csusb.edu

Tareq Nasralah
Northeastern University
t.nasralah@northeastern.edu

***Abstract*—In this study, we aim to analyze the public perception of Twitter users with respect to the use of ChatGPT and the potential bias in its responses. Sentiment and emotion analysis were also analyzed. Analysis of 5,962 English tweets showed that Twitter users were concerned about six main types of biases, namely: political, ideological, data & algorithmic, gender, racial, cultural, and confirmation biases. Sentiment analysis showed that most of the users reflected a neutral sentiment, followed by negative and positive sentiment. Emotion analysis mainly reflected anger, disgust, and sadness with respect to bias concerns with ChatGPT use.**

Keywords: ChatGPT, bias, social media, analytics, pre-trained model.

I. INTRODUCTION

The first few months of 2023 have seen a rise of artificial intelligence (AI) based large language models (LLM) such as ChatGPT- by Open AI. ChatGPT has seen a record breaking number of users in the first few months of its existence [1] and becomes a widely used tools for general question-answering and information seeking [2], [3]. This is made possible through ChatGPT's ability in finding and detecting patterns in a large dataset and the ability to improve performance through feedback [4].

With the widespread use of LLMs and Generative AI in different domains, AI bias is becoming problematic and more apparent [5]. Bias in the context of LLM could be defined as the "presence of systematic misrepresentations, attribution errors, or factual distortions that result in favoring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions based on learned patterns" [6]. Although LLM, such as GPT, tend to claim they are not impartial, some research suggests that they suffer from race, gender, religion, and political orientation bias [7]. Such bias increased the awareness of the impact these LLM might have on people in general [8].

The objective of this study is to analyze the public perceptions of bias present in ChatGPT by analyzing social media posts on Twitter using machine learning techniques. More specifically, we plan to identify and categorize the types of bias that are currently present in ChatGPT as expressed on Twitter. Identification of these biases could help researchers develop new mitigation strategies for biases in AI and help better determine a set of socio-ethical and legal principles that need to be considered when implementing similar AI systems [5].

II. LITERATURE REVIEW

The recent popularity of ChatGPT is driving significant research on the topic. Motoki et al [7] investigated the political bias concerns posed by ChatGPT following an empirical design by asking the new LLM to impersonate someone from a given side of the political spectrum and ask a number of questions with already known answers. The authors used several answers for the same question and completely randomized the order of questions being asked to ChatGPT over several rounds. Results showed that ChatGPT has a significant and systematic bias towards a specific political party in different countries.

Praveen & Vajrobol [9] used bidirectional encoder representations from the transformers model to study healthcare researchers' perceptions about ChatGPT. Using 62,734 tweets obtained with a search query consisting of the word 'ChatGPT' and a phrase that helps identify the users who posted the tweets working as a healthcare researcher, such as 'I am a healthcare researcher.' BERT was used to analyze tweets with respect to sentients and topics being discussed, using topic modeling. Results showed that 51.4% of the tweets

have neutral sentiments, 33.7% of the tweets have positive sentiments, and 14.7% of the tweets have negative sentiments. Topics identified were mainly related to ChatGPT being helpful in research in general, promoting big data analytics, assisting in reading research papers, and researchers being doubtful about the accuracy of the model.

Taecharungroj [10] analyzed Twitter data to determine what ChatGPT can do for users. The authors collected a total of 233,914 English tweets about ChatGPT. Tweets were analyzed using LDA topic modeling and thematic analysis. Topic modeling and thematic analysis revealed three main topics. These topics were related to news, technology, and reactions. In addition, five functional domains were also identified. These functional domains were related to "creative writing, essay writing, prompt writing, code writing, and answering questions."

Leiter et al. [11] analyzed Twitter users' perceptions of ChatGPT. The authors analyzed the sentiments of the tweets, the change of sentiments over time, sentiments across languages, as well as analysis of topics. Using a simple search query that consists of '#ChatGPT,' the authors collected over 330,000 tweets from over 168,000 Twitter users. Results and analysis showed that 100,163 tweets have positive sentiment, 174,684 tweets have neutral sentiment, and 59,961 tweets have negative sentiment. Sentiment analysis over time showed a downward trend of sentiment during the analyzed time frame. The average sentiment of English vs. non-English tweets was pretty much similar. Tweets in English have more positive sentiments compared to tweets in Japanese, French, Spanish, and German. Finally, tweets were mainly related to five topics: business, technology, education, daily life, and social concerns.

Korkmaz et al. [12] studied the sentiments of those who experience ChatGPT using social media analysis. Using a simple search query, 'ChatGPT,' the authors collected about 788,000 English tweets. Tweets were analyzed using AFINN, Bing, and NRC sentiment dictionaries. Analysis of the results from the three sentiment dictionaries showed that Twitter users were satisfied with ChatGPT and found the experience using the LLM successful. On the other hand, many users reported negative sentiments and emotions, such as fear and concern with respect to ChatGPT.

The review of the literature showed that few studies had addressed ChatGPT using social media data [9]–[13]. Furthermore, despite the fact that the topic of bias and AI has been widely studied in the literature [5], [14]–[22], there is a very limited body of literature that addressed bias in emerging technology like ChatGPT [6], [7], [23]. Accordingly, this study attempts to address the public perception of ChatGPT bias by analyzing social media data from Twitter. The study also aims at analyzing sentiments, emotions, and volumes of tweets across different types of biases.

III. Research Design and Methodology

Figure 1 shows a high-level overview of the research methodology followed to identify and analyze the types of ChatGPT related biases discussed by the public on the Twitter social media platform.

Twitter was used to collect public tweets about ChatGPT bias using a simple search query (ChatGPT and Bias) of English tweets between November 1st, 2022, and May 31st, 2023. The collected tweets were analyzed for sentiments. Sentiment analysis is a well-known area of NLP used to determine the type of sentiment polarity (positive, neutral, and negative) from the text [24], [25]. In addition, emotion analysis utilizing the "Ekman 6" (Anger, Fear, Disgust, Joy, Surprise, and Sadness) [26] was also completed.

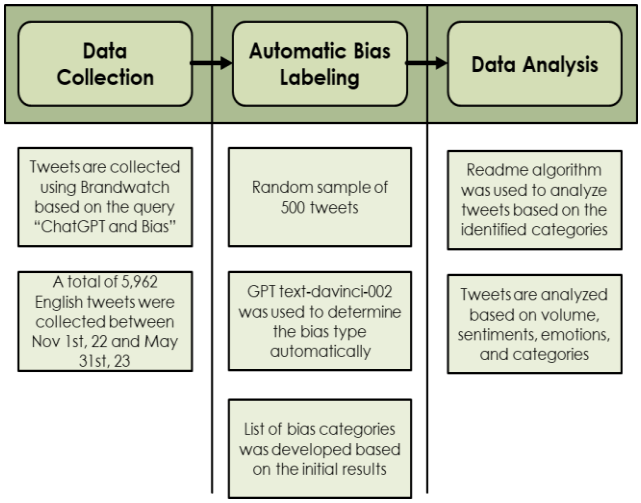

Figure 1. Research methodology for identifying ChatGPT bias types.

A random sample of 500 tweets was selected as input for the text-davinci-002 GPT model to automatically determine the type of bias being discussed by the public. The text-davinci-002 is an "InstructGPT model which utilizes Reinforcement Learning with reward models trained based on human comparisons" [27]. The text-davinci-002 was used through OpenAI's API and a Python code that requires prompt specifications. The prompt for the task on hand was "What is the type of bias mentioned in the text? Provide the type of bias without any explanation. If the text does not explicitly state the type of bias, then do not try to come up with an approximation for the bias type and simply state 'Unknown Bias'".

To validate the results by the text-davinci-002 model, out of the 500 labeled tweets, a random sample of 60 tweets was manually labeled, by one of the authors, with the bias type and then compared with results from the model. Inter-rater reliability [28] was calculated to avoid any bias in the results from the model and one of the authors.

The results from the automatic labeling were grouped into a higher level of bias categories. The categories were then used to help with the automatic labeling of the remaining tweets. To do so, a custom classifier was created in Brandwatch using the ReadMe algorithm developed by Hopkins & King (2010). The algorithm emphasizes social science goals, focusing on a broad categorization of the whole sets of documents (Hopkins & King, 2010) and showing how tweets spread across the different types of biases and give an unbiased text classification compared to traditional supervised learning techniques.

Based on the results obtained from the ReadMe classifier, we have analyzed the distribution of tweets across different types of biases and performed emotions and sentiment analysis.

IV. RESULTS

We collected a total of 5,962 tweets posted by 5,235 users. Among those who stated their gender, 418 (16%) were females and 2118 (84%) were males. Figure 2 shows the volume of tweets over time.

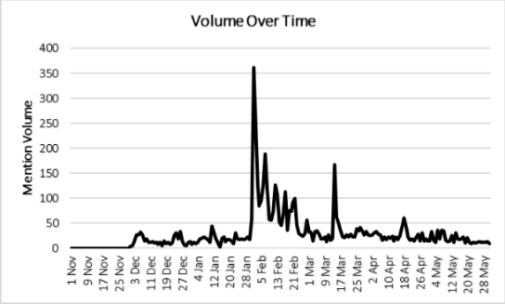

Figure 2. Tweets volume over time.

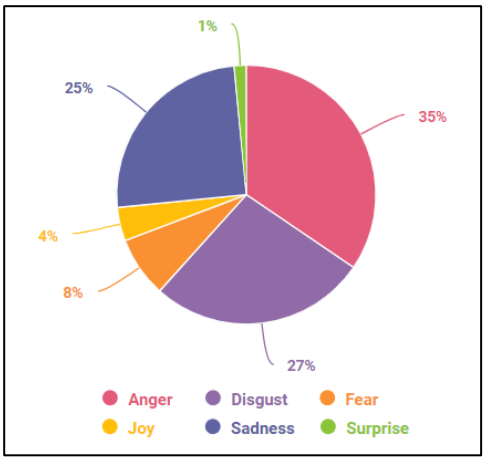

Figure 3. Emotion analysis

The emotion analysis shown in Figure 3. The figure shows that 35% of the posts reflected anger emotion, which indicates users' outrage against bias in ChatGPT, followed by 27% of the posts reflecting disgust and 25% of the posts reflecting sadness.

Figure 4 summarizes the results of the sentiment analysis where 64% of the posts reflect neutral sentiment towards bias in ChatGPT, 34% of the posts reflect negative sentiment towards bias in ChatGPT, and only 2% reflect positive sentiment.

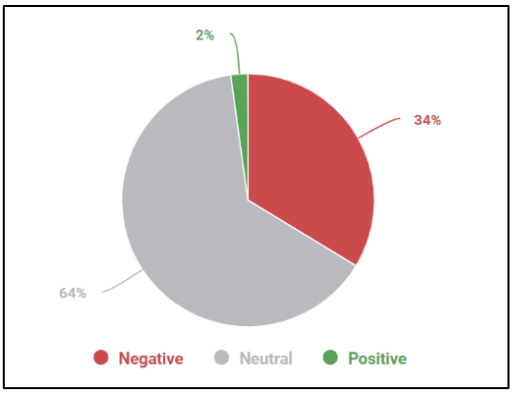

Figure 4. Sentiment analysis

Figure 5 shows the analysis of ChatGPT bias types as reflected in the public's tweets. The separate manual qualitative analysis for the sample of 60 tweets results in Cohen's Kappa statistics of 87%, which reflects almost perfect agreement among the researcher and the results from the text-davinci-002 model [28].

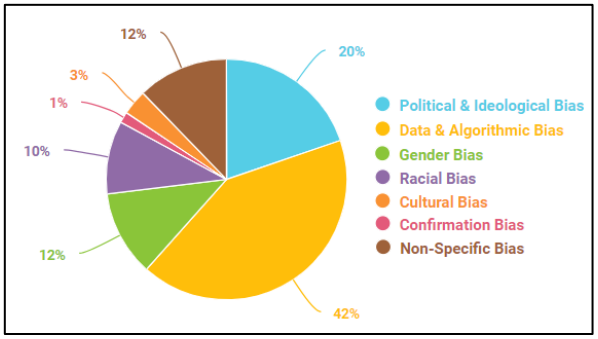

Figure 5. ChatGPT bias types

Most of the tweets (42%) discussed data and algorithmic bias-related followed by political and ideological bias (20%), socio-cultural bias, including gender bias (12%), racial bias (10%), and cultural bias (3%). Additionally, (1%) of the tweets mentioned confirmation bias. The category of 'Non-Specific Bias' (12%) represents tweets that discussed bias in general without specifying a particular type.

Figure 6 shows the volume of tweets by the Twitter users' profession. The top three professions concerned about ChatGPT bias, among those who stated their professions on Twitter, were executive (391 tweets – 6.6%), software developer and IT (238 tweets - 4%), and scientists and researchers (230 tweets - 3.9%). Other professions have less than 200 tweets each.

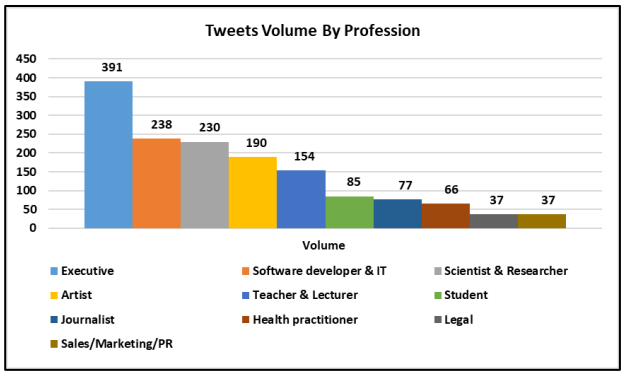

Figure 6. Tweets volume by profession

Figure 7 shows the volume of tweets by the Twitter users' interests. The top three categories of users' interests who were concerned about ChatGPT were technology (885 tweets - 14.8%), business (662 tweets - 11.1%), and politics (546 tweets - 9.2%). Other interests have less than 500 tweets each.

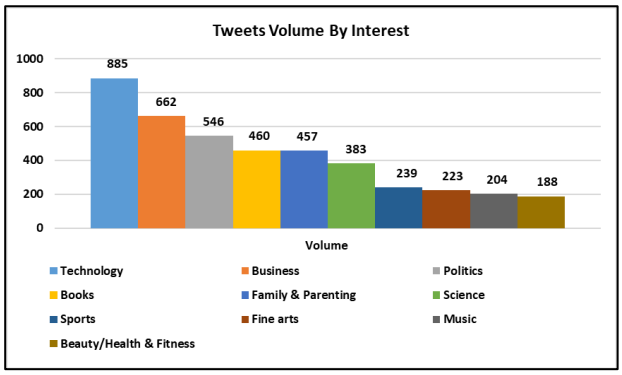

Figure 7. Tweets volume by interest

V. DISCUSSION

The study of ChatGPT bias through the analysis of 5,962 English tweets showed that Twitter users were mainly concerned about data and algorithmic bias, political and ideological bias, gender bias, racial bias, and cultural bias.

Data and algorithmic bias were the main concern expressed by Twitter users. These biases could be attributed to several factors, as expressed by Twitter users, including biases originating from ChatGPT's design, training data, algorithmic functions, ChatGPT's built-in biases, and computational bias. Additionally, ChatGPT may inherit biases from its developers and the data it is trained on. Example tweet expressing concerns about data and algorithmic bias:

> "*Bias in AI programming needs to be legislated against in equalities legislation to tackle companies like Microsoft programming ChatGPT with definitive religious bias in AI programming*", "*Regulating smart tech platforms like #ChatGPT is necessary to address algorithmic bias and copyright concerns*", and "*And there it is, programmer bias. How intelligent is ChatGPT really, if can't distinguish between what's real and the inherent political bias in it's code?*"

According to the literature, algorithmic bias, often referred to as the main contributor to various types of AI bias, could be observed in many AI systems [30]. Algorithmic bias could arise from biases in the modeling process or biases during training [5] as AI models are typically trained on large datasets. However, little attention is paid by developers to the collection and processing of such data [22]. As a result, AI algorithms are frequently trained on biased datasets, which could negatively affect the quality of the AI model [31].

Political and ideological bias emerged as another major concern among Twitter users. Users highlighted various types of bias contributing to this concern, including left-wing bias, right-wing bias, liberal bias, conservative bias, and biases associated with specific ideologies and affiliations. Example tweet expressing concerns about political and ideological bias:

> "*ChatGPT is hard left-green fascist. It is dangerous. The political bias is extreme*", "*The political bias was built into the dataset before ChatGPT became operational*", "*It is not coincidence that ChatGPT holds same political bias, passive aggressive language of its BigTech-Liberal Overlords*", "*I'm definitely not using that. left wing bias is baked into ChatGPT*", and "*ChatGPT can't be trusted fully to be accurate as it has a political/ideological bias*."

This type of bias aligns with existing literature about AI and political bias. Political bias could manifest when an AI system demonstrates a preference for certain individuals, groups, or content based on their political orientation [30]. In a recent study conducted by Motoki et al. (2023) on ChatGPT and political bias, the results revealed a significant and systematic bias towards a specific political party across different countries.

Socio-cultural bias consists of gender bias, racial bias, and cultural bias. When developing products and systems, we aim to promote diversity, equity, and inclusion among various user groups. However, achieving this goal is often hindered by incorrect training data, weak algorithm design, and, most importantly, deep-rooted socio-cultural biases [32]. Users reported several reasons behind the presence of socio-cultural bias in ChatGPT, including cultural or geographic norms, values, or viewpoints, gender bias, racial bias, religious bias, and biases against specific racial or ethnic groups.

Simple AI applications have encountered certain limitations during their early stages. For example, Google Translate used to refer to women as 'he said' or 'he wrote' when translating from Spanish to English. Another example involves camera software that wrongly interprets Asians as always blinking when warning against blinking in photos [22]. Furthermore, certain AI applications have demonstrated discriminatory behavior against specific groups of people [22]. For example, some AI based apps tend to characterize names of European people as pleasant while names of African names

as unpleasant [22]. According to Nadeem et al. (2020), AI bias related to race and gender could be attributed to factors such as the lack of diversity in the "data and developers, the bias in society, and bias in data due to programmer conscious or unconscious bias" [33]. Example tweet expressing concerns about sociocultural bias:

> "*There's clearly some racial as well as gender bias within ChatGPT and it really calls into question the integrity of AI technology*", "*Apparently ChatGPT has gender bias*", "*I am amazed at how racist and bias ChatGPT is. Big tech is trying to force us all into a bubble*", and "*Interesting seeing some of the cultural bias already manifesting in how these AIs are trained though, just because a dataset points in a certain way does not mean it's right. Saw this when prompting on vague topics, when getting more specific, things tended to even out #ChatGPT*"

Confirmation bias has been another concern expressed by Twitter users. Confirmation bias "connotes the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" [34]. Example tweet expressing concerns about confirmation bias:

> "*ChatGPT is a Confirmation Bias Machine. It will happily affirm with facts and reasoning anything you already believe.*", "*ChatGPT is too prone to confirmation bias, so bigots can influence it a lot. So it's useful for those things that are not subject to emotions.*", and "*It should be able to take more adversarial stances. I've dealt with a lot of confirmation bias when talking with ChatGPT.*"

According to Schwartz et al. [35], when it comes to AI projects, solid experimental design and the need for minimizing confirmation bias are being downplayed by many developers. AI confirmation bias usually occurs when the AI model is dealing with patterns in the data that are already known to the AI model, even if they are erroneous [36].

Sentiment analysis revealed that most of the tweets had a neutral sentiment followed by negative and positive sentiments. These results are consistent with existing research on AI and sentiments. For example, Praveen & Vajrobol [9] reported that 51.4% of the analyzed tweets exhibited a neutral sentiment. However, positive sentiment (33.7% of tweets) was more prevalent than negative sentiment (14.7%). On the other hand, Leiter et al. [11] found that the majority of the analyzed tweets revealed a negative sentiment (52.2%), followed by positive sentiment (29.9%) and neutral sentiment (19.9%). With respect to emotion analysis, most of the tweets reflected anger (35%), disgust (27%), sadness (25%), fear (8%), joy (4%), and surprise (1%). These results show consistency with respect to bias in ChatGPT.

Finally, volume analysis of tweets based on Twitter users' professions and interests revealed that the top three professions by volume are executives, software developers & IT, and scientists & researchers. Similarly, the top three interests in terms of volume were technology, business, and politics. These findings align with the analysis of different types of biases as well as the corresponding discussions, where the two most prominent categories of biases were data and algorithmic bias, as well as political and ideological bias.

VI. Conclusion

This study analyzed the public perception of ChatGPT bias by analyzing Twitter data using different analytical techniques. Results and data analysis revealed that different types of biases were reported by Twitter users. These biases were related to data and algorithmic bias, political and ideological bias, gender bias, racial bias, and cultural bias. Furthermore, sentiment and emotion analysis showed that most Twitter users reported neutral sentiment about ChatGPT, followed by negative sentiment and positive sentiment. The top three types of emotions that Twitter users reflected were: anger, disgust, and sadness about ChatGPT. Sentiment and emotion analysis reflected a negative user experience or thoughts about ChatGPT.

This work is not without any limitations. The query retrieved tweets that mentioned both "ChatGPT" and "Bias." To handle tweets that discussed multiple types of biases or did not explicitly mention the type of bias, we utilized the custom ReadMe classifier. However, it should be noted that the collected tweets do not provide direct evidence of users personally experiencing these types of biases when using ChatGPT. Instead, they primarily reflect the public opinion about ChatGPT bias. Additionally, the identification of bias types relied on a GPT model, which, while achieving a high accuracy rate as demonstrated by the inter-rater reliability results, is not fully accurate.

References

[1] J. M. Buriak *et al.*, "Best Practices for Using AI When Writing Scientific Manuscripts: Caution, Care, and Consideration: Creative Science Depends on It," *ACS nano*, vol. 17, no. 5, pp. 4091–4093, 2023.

[2] S. Ghosh and A. Caliskan, "ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages," 2023.

[3] N. Grant, "Google Calls In Help From Larry Page and Sergey Brin for AI Fight," *The New York Times*, *online*, 2023.

[4] N. Lee, P. Resnick, and G. Barton, "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms," *Brookings*, May 22, 2019. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ (accessed Jun. 05, 2023).

[5] X. Ferrer, T. van Nuenen, J. M. Such, M. Coté, and N. Criado, "Bias and Discrimination in AI: A Cross-Disciplinary Perspective," *IEEE Technology and Society Magazine*, vol. 40, no. 2, pp. 72–80, Jun. 2021, doi: 10.1109/MTS.2021.3056293.

[6] E. Ferrara, "Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models." arXiv, Apr. 18, 2023. Accessed: Jun. 03, 2023. [Online]. Available: http://arxiv.org/abs/2304.03738

[7] F. Motoki, V. Pinho Neto, and V. Rodrigues, "More Human than Human: Measuring ChatGPT Political Bias." Rochester, NY, Mar. 17, 2023. doi: 10.2139/ssrn.4372349.

[8] D. Roselli, J. Matthews, and N. Talagala, "Managing Bias in AI," in *Companion Proceedings of The 2019 World Wide Web Conference*,

San Francisco USA: ACM, May 2019, pp. 539–544. doi: 10.1145/3308560.3317590.

[9] S. V. Praveen and V. Vajrobol, "Understanding the Perceptions of Healthcare Researchers Regarding ChatGPT: A Study Based on Bidirectional Encoder Representation from Transformers (BERT) Sentiment Analysis and Topic Modeling," *Ann Biomed Eng*, May 2023, doi: 10.1007/s10439-023-03222-0.

[10] V. Taecharungroj, "'What Can ChatGPT Do?' Analyzing Early Reactions to the Innovative AI Chatbot on Twitter," *BDCC*, vol. 7, no. 1, p. 35, Feb. 2023, doi: 10.3390/bdcc7010035.

[11] C. Leiter *et al.*, "ChatGPT: A Meta-Analysis after 2.5 Months." arXiv, Feb. 20, 2023. Accessed: May 22, 2023. [Online]. Available: http://arxiv.org/abs/2302.13795

[12] A. Korkmaz, C. Aktürk, and T. TALAN, "Analyzing the User's Sentiments of ChatGPT Using Twitter Data," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 2, pp. 202–214, 2023.

[13] M. U. Haque, I. Dharmadasa, Z. T. Sworna, R. N. Rajapakse, and H. Ahmad, "'I think this is the most disruptive technology': Exploring Sentiments of ChatGPT Early Adopters using Twitter Data." arXiv, Dec. 12, 2022. Accessed: May 22, 2023. [Online]. Available: http://arxiv.org/abs/2212.05856

[14] S. Kapur, "Reducing racial bias in AI models for clinical use requires a top-down intervention," *Nat Mach Intell*, vol. 3, no. 6, Art. no. 6, Jun. 2021, doi: 10.1038/s42256-021-00362-7.

[15] T. C. Moran, "Racial technological bias and the white, feminine voice of AI VAs," *Communication and Critical/Cultural Studies*, vol. 18, no. 1, pp. 19–36, 2021.

[16] A. Nadeem, O. Marjanovic, and B. Abedin, "Gender Bias in AI: Implications for Managerial Practices," in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, D. Dennehy, A. Griva, N. Pouloudi, Y. K. Dwivedi, I. Pappas, and M. Mäntymäki, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 259–270. doi: 10.1007/978-3-030-85447-8_23.

[17] A. Nadeem, O. Marjanovic, and B. Abedin, "Gender bias in AI-based decision-making systems: a systematic literature review," *Australasian Journal of Information Systems*, vol. 26, 2022.

[18] S. O'Connor and H. Liu, "Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities," *AI & Soc*, May 2023, doi: 10.1007/s00146-023-01675-4.

[19] A. Pena, I. Serna, A. Morales, and J. Fierrez, "Bias in multimodal AI: Testbed for fair automatic recruitment," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 28–29.

[20] A. H. Sham *et al.*, "Ethical AI in facial expression analysis: racial bias," *SIViP*, vol. 17, no. 2, pp. 399–406, Mar. 2023, doi: 10.1007/s11760-022-02246-8.

[21] F. Sperrle, U. Schlegel, M. El-Assady, and D. Keim, "Human trust modeling for bias mitigation in artificial intelligence," presented at the ACM CHI 2019 Workshop: Where is the Human? Bridging the Gap Between AI and HCI, 2019.

[22] J. Zou and L. Schiebinger, "AI can be sexist and racist — it's time to make it fair," *Nature*, vol. 559, no. 7714, pp. 324–326, Jul. 2018, doi: 10.1038/d41586-018-05707-8.

[23] J. M. Toro, "Emergence of a phonological bias in ChatGPT," *arXiv preprint arXiv:2305.15929*, 2023.

[24] S. Mansour, "Social Media Analysis of User's Responses to Terrorism Using Sentiment Analysis and Text Mining," *Procedia Computer Science*, vol. 140, pp. 95–103, 2018, doi: 10.1016/j.procs.2018.10.297.

[25] A. Wahbeh, T. Nasralah, O. El-Gayar, M. Al-Ramahi, and A. Elnoshokaty, "Adverse Health Effects of Kratom: An Analysis of Social Media Data," in *Hawaii International Conference on System Sciences*, 2021. doi: 10.24251/HICSS.2021.477.

[26] P. Ekman, "Facial expression and emotion.," *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[27] A. Hendy *et al.*, "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation." arXiv, Feb. 17, 2023. Accessed: Jun. 05, 2023. [Online]. Available: http://arxiv.org/abs/2302.09210

[28] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data.," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977, doi: 10.2307/2529310.

[29] D. Hopkins and G. King, "A method of automated nonparametric content analysis for social science," *American Journal of Political Science*, vol. 54, no. 1, pp. 229–247, 2010.

[30] U. Peters, "Algorithmic Political Bias in Artificial Intelligence Systems," *Philos. Technol.*, vol. 35, no. 2, p. 25, Mar. 2022, doi: 10.1007/s13347-022-00512-8.

[31] S. 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*, in GE '18. New York, NY, USA: Association for Computing Machinery, May 2018, pp. 14–16. doi: 10.1145/3195570.3195580.

[32] S. Akter, Y. K. Dwivedi, S. Sajib, K. Biswas, R. J. Bandara, and K. Michael, "Algorithmic bias in machine learning-based marketing models," *Journal of Business Research*, vol. 144, pp. 201–216, 2022.

[33] A. Nadeem, B. Abedin, and O. Marjanovic, "Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies," *ACIS 2020 Proceedings*, Jan. 2020, [Online]. Available: https://aisel.aisnet.org/acis2020/27

[34] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of general psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[35] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a standard for identifying and managing bias in artificial intelligence," *NIST Special Publication*, vol. 1270, pp. 1–77, 2022.

[36] B. Mueller, T. Kinoshita, A. Peebles, M. A. Graber, and S. Lee, "Artificial intelligence and machine learning in emergency medicine: a narrative review," *Acute medicine & surgery*, vol. 9, no. 1, p. e740, 2022.