Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The Ethics of AI-Generated Content: Combating Bias and Misinformation in Generative Models

Arifa Khan[1], Dr.P. Saravanan[2*]
[1] *Research Scholar, Management, SRM Institute Science of Technology, SRM Nagar, Katankulattur, Tamil Nadu, India.*
ak7641@srmist.edu.in
[2] *Associate Professor, Management, SRM Institute Science of Technology, SRM Nagar, Katankulattur , Tamil Nadu, India.*
saravanp2@srmist.edu.in

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

The development in AI has made it really easy for various models to generate contents like image, sound and text. AIGC have transformed the way in which various domains from media to entertainment develop and create content. The output generated from the AI model is highly realistic, leading to questions on whether it can be trusted or not. AI-generated media content has given rise to ethical issues related to bias and information. It has become important to take appropriate steps to mitigate these ethical issues for harnessing the potential of AI and minimizing the impact. This paper explores the methods and techniques that can be used to address the two ethical issues of bias and misinformation. Secondary information collected from online articles and peer-reviewed journals have been incorporated to provide the necessary discussion on the subject. The current methods and approaches have effectively been discussed in the paper which can help address the concerns and generate trustworthy content.

Keywords: AIGC, Ethical, Misinformation, Bias, Models, Algorithms

INTRODUCTION

We are living in a digital world that is witnessing technological advancements so quickly that it is difficult to track the progress in the field. The field that has made the most progress in the last few years is artificial intelligence (AI), emerging as a revolutionary force. It has affected multiple facets of human existence and become an essential part of our daily lives (Nader et al. 2022). The importance of AI in our lives is only going to increase with time. The application of AI can be visible in all segments including healthcare, banking, transportation, entertainment and media (Arrieta et al. 2020). The most recent development in AI is the way it has made it easy to generate media content like images, texts and sounds. AI has made it easy to create images that imitate the way we look, write texts that way we write and generate sound in a way that we speak (Partadiredja et al. 2020). The content generated with the help of AI is quite appealing and looks highly realistic, making it difficult to tell the difference between AI and human. The growth of AI-generated content (AIGC) has led to various discussions over the ethical aspects of these contents. Generative AI models like “Deepfakes” have opened a whole new world, enabling the generation and manipulation of digital contents based on specific instructions (Xu et al. 2023). There are significant concerns related to these contents especially around misinformation and bias.

The AIGCs like text, images and videos have increased the potential for disseminating false information that can have a significant impact on social integrity and individual rights. The spread of misinformation using AIGC can have a significant impact on health, public opinion and democracy (Bashardoust et al. 2024). AI can generate content that can quite convincingly communicate false narrative and make it difficult to fight misinformation. On the other hand, there is an issue of bias with these contents which can come from algorithm design, training data and implementation processes. The bias can result in the reinforcement of stereotypes and discrimination, undermining all the efforts that are being undertaken for creating a fair and inclusive society (Marinucci et al. 2023). Hence, this research explores the ethical aspects of AIGC by focusing on two of the key challenges- bias and information. The paper will discuss the techniques and methods for combating bias along with the methods and approaches of addressing misinformation.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

METHODOLOGY

The ethical concern around AIGCs has been the topic of interest for researchers across the globe. There has been a significant amount of research done on the subject and thus this research will make use of secondary data to address the goals of the research. Secondary sources refers to the use of information and data that has been published by other researchers or institutions. It makes it easy to collect a large amount of data on the topic and gain a comprehensive idea on the topic being explored. Hence, information from peer-reviewed journals and online articles has been used to develop this paper and provide the required insights on the topic.

FINDINGS AND DISCUSSION

3.1 Application of AI-generated contents in different domains

The contents generated using AI have found their adoption in various domains because of their efficiency and high-quality outputs. It has emerged as a transformative force that is able to produce content that is tailored to the needs of the individual or institution. The content is able to engage with the audience in the most effective manner and pushes the boundary of creativity and innovation. The domains that have witnessed significant application of such contents are:

Marketing and advertising- Generative AI or Gen AI has completely transformed the way in which brands interact with their customers (Lim et al. 2024). Gen AI models like Chat GPT and DALL-E have become quite popular and are being adopted by marketers for better ad targeting. A survey conducted in North America, South America and Europe involving 202 marketing and advertising professionals found that 75% of them used Chat GPT for generating content (Dencheva, 2024). AI has made it easy to generate personalized messages to target specific customers in order to improve conversion rates and increase engagement. GPT-4 can easily be used to create blog posts, write product descriptions and develop social media content in a consistent manner while saving both time and money.

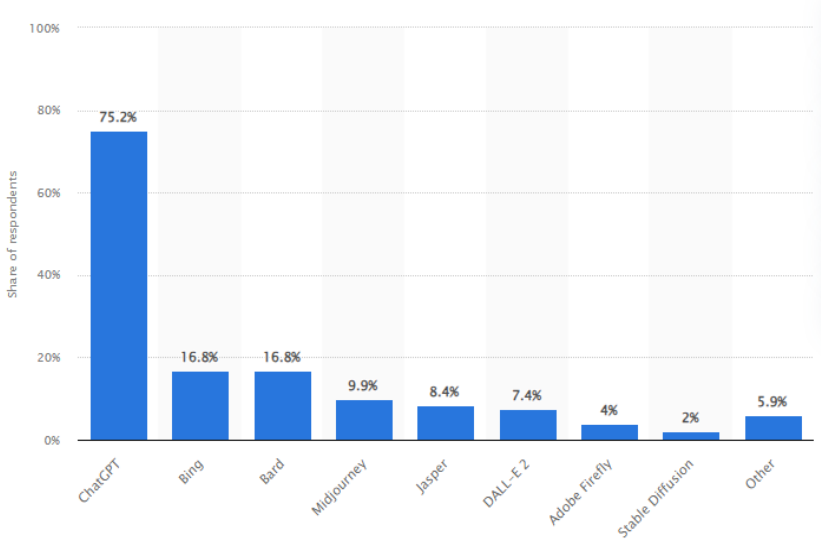

Figure 1: Gen AI tools used for marketing in 2023
(Source: Dencheva, 2024)

Journalism and media- Gen AI has emerged as a practical tool that is widely being used in the world of journalism and media. The top media houses like NY Times, Associated Press, Bloomberg and Forbes have deployed the capabilities of AI in their news business (Zagorulko, 2023). AI is being used in this domain for data extraction, processing and analysis, fact-checking, news report generation and speech recognition. There is numerous evidence of media outlets and journalists using AI to generate a particular content. These contents are changing the world of journalism by improving the speed and breadth of news coverage. Forbes uses an AI tool named Bertie for developing first drafts, headlines and topics for writers while Bloomberg uses Cyborg technology for generating thousands of articles related to financial news (Saad, 2020). The Washington Post also has an in-house AI tool named Heliograf for generating short reports on election results, sports and other important stories (Saad, 2020). Hence, AIGC in this

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

particular domain has not only improved efficiency and scalability but also opens new opportunities for storytelling and audience engagement.

Entertainment and creative arts- The scope of application of AI in this industry is growing significantly with the ability of AI to create new forms of content while enhancing existing ones (Amato et al. 2019). There is numerous evidence of AI tools being used for generating plot ideas, composing music, character interactions, generating artworks and producing entire albums. These models have made it easy to replicate the voice of everyone from Taylor Swift to Donald Trump with numerous evidence of deepfakes versions available on the internet which contain speeches that were never made by these celebrities. On the other hand, there are positive examples of AI being used in the industry like Val Kilmer using the technology to recreate his voice for Top Gun: Maverick. Similarly, the biggest music label in South Korea HYBE used AI to release a song by an artist in six different languages (Li & Bantourakis, 2023). AI models are also trained to generate scripts, storyline, voice recording and editing. The application of AI is only going to increase in this industry with time and at various levels with large amounts of content being generated with the help of these tools.

Marketing & advertising, journalism & media and entertainment & creative arts are the three main domains that are using AIGC in a significant manner and the applications of AI contents in these domains are only going to increase with time. Other domains like education, health, finance and legal also used such contents but the consumption of AI contents is quite high in these three domains. One of the primary reasons for using AI content is cost savings by industries. A survey of 1000 business leaders in the US found that 24% of the companies were able to save 50,000 to 70,000 dollars with the help of GPT with others also saving a reasonable amount of money (Thormundsson, 2023).

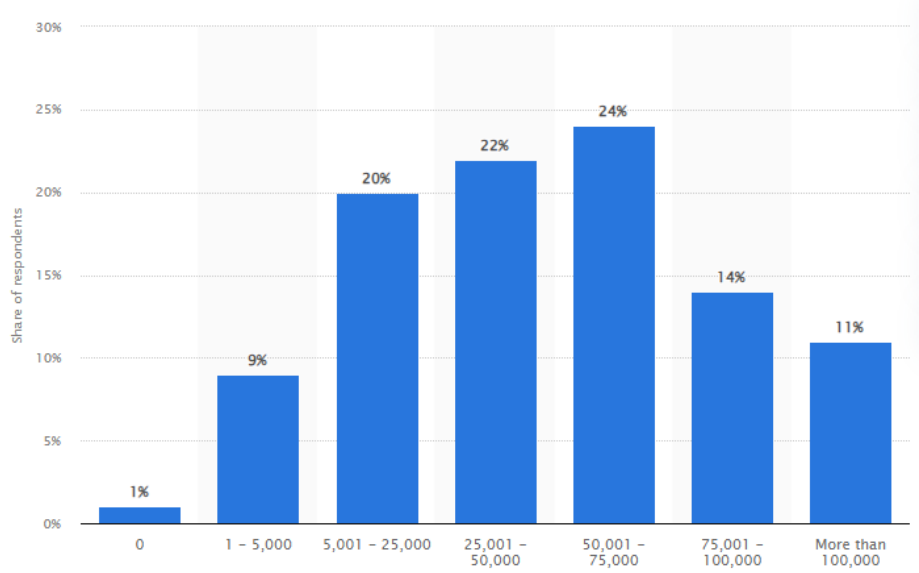

Figure 2: Money saved by US businesses by using GPT
(Source: Thormundsson, 2023)

Despite the potential of AI contents in various domains, there are various ethical issues like misinformation and bias that arise related to the contents and thus the same have been discussed in the next section of the paper.

3.2 Ethical concerns in AI-generated content

The two most prominent ethical concerns with AIGC are bias and misinformation. Bias can be present in the content from various sources and in turn influence the outcome. Large scale AI models like Chat GPT and LLaMA are trained with the use of massive data to understand human languages (Ouyang et al. 2020). These trained AI models then generate content based on the prompts provided by the users which in turn is referred to as the AIGC. However, there can be biases present in the training data that can impact the content generated along the way. The massive amount of data that is used to train the AI models may contain many historical biases related to gender, race and culture (Ntoutsi et al. 2020). The presence of such biases in the data would lead to AI replicating the same biases in its

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

outputs. For example, if the training data has more positive things to say about male professionals than female professionals then the content generated would be male oriented rather than being neutral or unbiased. Similarly, the bias can also arise from the selection of data that has been used to train the AI models. If a particular group is underrepresented or overrepresented in the training data then the results produced would be biased. For example, an AI model that has been trained using Western literature is more likely to generate contents that do not favor non-western literature. On the other hand, the design of the AI model can in itself lead to bias with some algorithms favoring certain types of data or patterns (Nazer et al. 2023). The criteria set for evaluating the model can also lead to bias. If the model is not designed or evaluated for fairness then the results generated would be biased despite the training data being balanced.

Misinformation in AIGC is another important ethical concern that is widespread with potentially harmful consequences. Misinformation, defined as false or misleading information, can be generated at a large scale with the help of AI models (Kreps et al. 2022). Malicious actors can use AI models to generate and circulate credible sounding news or stories at a large scale. In context to AIGC, misinformation can occur due to the errors present in the training data or the inability of the model to distinguish between true or false information. As mentioned earlier, AI models are trained over a large dataset but if the dataset contains false information, outdated content and other inaccuracies, then the content can lead to these particular errors. AI models do not have the capability to verify the credibility of the data and generate content by identifying patterns without the understanding of the truth (Brundage et al. 2020). On the other hand, misleading or wrong information can be entered deliberately by malicious actors to generate wrong content. The spread of misinformation using AIGC can harm individuals or communities and cause political unrest while eroding public trust. It is very important to address these ethical concerns related to AIGC in order to harness its full potential and minimize the impacts.

3.3 Techniques and methods to combat bias in AIGC

The ethical use of AI that ensures fairness and transparency can only be done by addressing the bias involved with the models. There are various methods and techniques that can be used to identify, mitigate and prevent the bias in the models in the way of generating appropriate content. The training dataset has been identified as one of the key sources contributing to the bias with the AIGC. The application of appropriate data preprocessing techniques on the training datasets needs to be done to make it free from bias. Data augmentation can enhance the diversity of the training datasets, ensuring that there is a balance with the representation of all the groups in the dataset (Sharma et al. 2020). This can also be done by oversampling the groups that are underrepresented and under-sampling the ones that are over represented. Data cleaning is another technique that can be used for removing biases and enhancing the quality of training data (Tae et al. 2019). Further, there are different metrics that can be used in bias detection by quantifying the bias along with systematic audits that can help in identifying the bias at each stage of the development. Apart from that, there are various tools that can be used for interpreting the AI models decisions and identifying the biases in the predictions. LIME (Local Interpretable Model-agnostic Explanations) is a widely popular tool that can be used to understand the behavior of the model in an appropriate manner (Zafar & Khan, 2021). Similarly, SHAP (SHapley Additive exPlanations) is a model that can help understand and interpret the output of the AI models (Antwarg et al. 2021). The aspect of bias can also be removed with the help of automated tools that scan and analyze the AI models to provide actionable insights.

With the presence of various techniques and methods for mitigating bias in the models, various studies have incorporated the same to provide appropriate insights into mitigating bias. One such research has been done by Fang et al. (2024) where data from 4860 articles from New York Times and 3769 articles from Reuters were collected to understand the extent to which the models were resistant to biased prompts. The AI models that were part of the research are ChatGPT, Cohere, and LLaMA in order to identify the bias in the AIGC from these models. The framework that was used for evaluating the bias has been provided in the below image. The investigation of all the models using the framework highlighted that ChatGPT has the lowest bias and this is generally because of the RLHF (reinforcement learning from human feedback) feature. The RLHF is effective in reducing the biases since it provides the ability for the model to decline content with biased prompts. RLHF is an important feature that can help mitigate bias and something that needs to be incorporated in all models. Similarly, the study also found that the bias of AIGC decreases with the increase in the size of the model.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

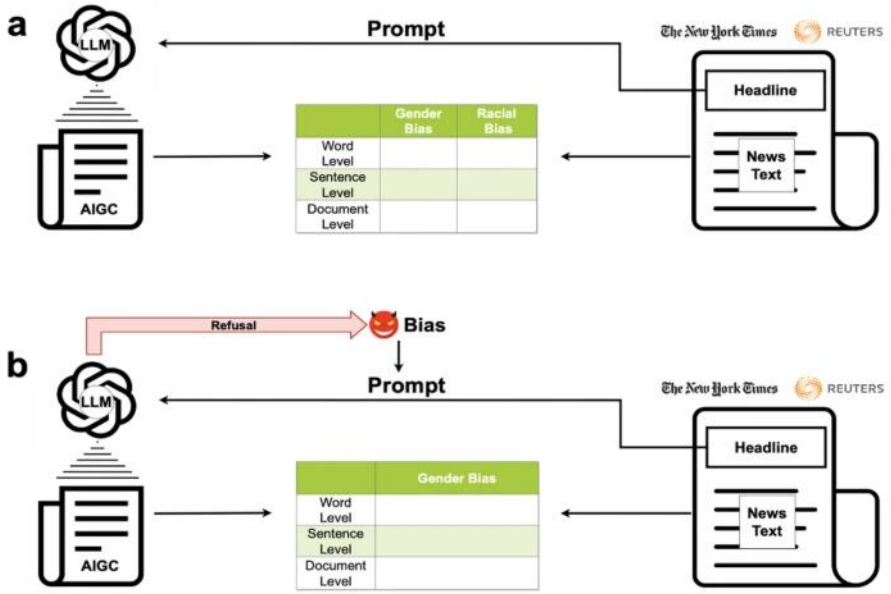

Figure 3: Framework for evaluating (a) Gender and racial bias of AIGC with unbiased prompts (b) Gender bias of AIGC under biased prompts
(Source: Fang et al. 2024)

The size of training data is only going to rise with time and they are going to become abundant, making it challenging to address bias yet very important. Hence, the different techniques and methods that can be used to address bias in AIGC as highlighted in the above discussion.

3.4 Methods and approaches for addressing misinformation

There are two approaches that have been explored to address the ethical issue of misinformation and they are- algorithm centered and human centered. The algorithmic approach involves the automatic detection and correction of misinformation along with its characterization while human-approach involves the way in which experts or people can help address misinformation (Zhou et al. 2023). Hence, both of these approaches for addressing misinformation related to AIGC have been discussed in this section of the article.

Algorithm-centered: The use of misinformation detection models has been found to reduce the belief of the people related to misinformation. NLP techniques can be used to check the generated content against facts and reputable sources. Generative Adversarial Networks (GANs) is another technique in this particular context where one network generates the content and the other network works at detecting the misinformation in the content (Hiriyannaiah et al. 2020). Further, algorithms that specifically focus on reducing the bias involved with the content can also be used in the process. The issue of misinformation can also be addressed by using algorithms to develop models that check for internal consistency. This approach focuses on developing numerous models that can address the issue. There is also a scope of developing models that can predict misinformation and generate contents that can counter the same. The issue can also be addressed by developing context aware models that can understand the context and situation to generate appropriate content and fight any misinformation. Hence, this particular approach equips the AI models to deal with misinformation in an effective manner and generate contents that are reliable and trustworthy.

Human-centered: This particular approach involves the involvement of experts and the general public to fight the issue of misinformation. The most common approach in this particular context is fact-checking which involves evaluating the credibility and correctness of the generated content through manual searchers (Zhou et al. 2023). This particular approach has been found to be quite effective in debunking misinformation but the labor intensive nature of the approach makes it difficult to scale. Another approach in this particular context is improving information literacy. Government institutions, scholars and journalists have developed guidelines and frameworks that can help the public spot misinformation. An educational approach has been found to be quite effective in fighting

misinformation and thus there is a need to develop an education agenda that takes into consideration the AI capabilities.

CONCLUSION

The impact of AI is only growing to increase across domains with vast amounts of content generated using the models. However, the large amount of AIGC also gives rise to ethical concerns related to misinformation and bias. The ethical issue of bias can be fought with the help of various methods like data augmentation, data cleaning, systematic audits, metrics, LIME, SHAP and RLHF. Similarly, the two approaches that can be used to address misinformation are algorithm-centered and human-centered. A detailed discussion into all of these elements have been done in the above research with the help of secondary sources. However, the research is not without its limitations and thus the use of secondary sources serves as one of the limitations. Another limitation in the paper is that the research has not included all the methods and approaches for addressing the ethical issues. There are still numerous methods or approaches that can be used to address the issue. Similarly, the paper has discussed only two of the ethical issues which are bias and misinformation while there are also other issues with AIGC such as accountability, transparency and integrity which have not been explored. Hence, there is significant scope to develop a comprehensive research paper in the future that explores all the ethical issues and outlines a myriad of methods that can be used to address the same. Despite the limitations, the research has been able to provide actionable insights for institutions and relevant stakeholders to address this ethical issue.

REFERENCES

[1] Partadiredja, R. A., Serrano, C. E., & Ljubenkov, D. (2020, November). AI or human: the socio-ethical implications of AI-generated media content. In *2020 13th CMI Conference on Cybersecurity and Privacy (CMI)-Digital Transformation-Potentials and Challenges (51275)* (pp. 1-6). IEEE. https://doi.org/10.1109/CMI51275.2020.9322673

[2] Lim, C. V., Zhu, Y. P., Omar, M., & Park, H. W. (2024). Decoding the Relationship of Artificial Intelligence, Advertising, and Generative Models. *Digital*, *4*(1), 244-270. https://doi.org/10.3390/digital4010013

[3] Dencheva, V., (2024) *Generative AI tools & platforms used in marketing & advertising worldwide 2023*. https://www.statista.com/statistics/1405052/gen-ai-tools-used-marketing-advertising/

[4] Zagorulko, D. I. (2023). ChatGPT in newsrooms: Adherence of AI-generated content to journalism standards and prospects for its implementation in digital media. *ВЧЕНІ ЗАПИСКИ*, *34*, 73. https://doi.org/10.32782/2710-4656/2023.1.2/50

[5] Li, C., & Bantourakis, M., (2023). *6 ways AI could disrupt the entertainment industry*. https://www.weforum.org/agenda/2023/08/hollywood-strike-synthetic-voice-digital-avatar-ai-entertainment/

[6] Amato, G., Behrmann, M., Bimbot, F., Caramiaux, B., Falchi, F., Garcia, A., & Vincent, E. (2019). AI in the media and creative industries. *arXiv preprint arXiv:1905.04175*. https://arxiv.org/pdf/1905.04175

[7] Thormundsson, B., (2023). *Amount of money companies in the United States saved by using ChatGPT as of February 2023*. https://www.statista.com/statistics/1379027/chatgpt-use-us-companies-money-saved/

[8] Saad Saad, D. T. A. (2020) Integration or Replacement: Journalism in the Era of Artificial Intelligence and Robot Journalism. *IJMJMC, 6*(33), 01-13. http://dx.doi.org/10.20431/2454-9479.0603001

[9] Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2024). Bias of AI-generated content: an examination of news produced by large language models. *Scientific Reports*, *14*(1), 1-20. https://doi.org/10.1038/s41598-024-55686-2

[10] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.

[11] Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of experimental political science*, *9*(1), 104-117. https://doi.org/10.1017/XPS.2020.37

[12] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356. https://doi.org/10.48550/arXiv.2008.07341

[13] Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS digital health*, *2*(6), e0000278. https://doi.org/10.1371/journal.pdig.0000278

[14] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*. https://doi.org/10.48550/arXiv.2004.07213

[15] Nader, K., Toprac, P., Scott, S. E., & Baker, S. W. (2022). Public understanding of artificial intelligence through entertainment media. *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01427-w

[16] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

[17] Marinucci, L., Mazzuca, C., & Gangemi, A. (2023). Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender. *AI & SOCIETY*, *38*(2), 747-761. https://doi.org/10.1007/s00146-022-01474-3

[18] Bashardoust, A., Feuerriegel, S., & Shrestha, Y. R. (2024). Comparing the willingness to share for human-generated vs. AI-generated fake news. *arXiv preprint arXiv:2402.07395*. https://doi.org/10.48550/arXiv.2402.07395

[19] Antwarg, L., Miller, R. M., Shapira, B., & Rokach, L. (2021). Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert systems with applications*, *186*, 115736. https://doi.org/10.1016/j.eswa.2021.115736

[20] Zafar, M. R., & Khan, N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, *3*(3), 525-541. https://doi.org/10.3390/make3030027

[21] Tae, K. H., Roh, Y., Oh, Y. H., Kim, H., & Whang, S. E. (2019). Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In *Proceedings of the 3rd international workshop on data management for end-to-end machine learning* (pp. 1-4). https://doi.org/10.1145/3329486.3329493

[22] Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D., Muthusamy, V., & Varshney, K. R. (2020). Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 358-364). https://doi.org/10.1145/3375627.3375865

[23] Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023, April). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-20). https://doi.org/10.1145/3544548.3581318

[24] Hiriyannaiah, S., Srinivas, A. M. D., Shetty, G. K., Siddesh, G. M., & Srinivasa, K. G. (2020). A computationally intelligent agent for detecting fake news using generative adversarial networks. In *Hybrid Computational Intelligence* (pp. 69-96). Academic Press. https://doi.org/10.1016/B978-0-12-818699-2.00004-4

[25] Xu, D., Fan, S., & Kankanhalli, M. (2023, October). Combating misinformation in the era of generative AI models. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 9291-9298). https://doi.org/10.1145/3581783.3612704