Analysis Effect of Bias lying in Artificial Intelligence: A Survey

Bangdao Liu [1], Keyao Wang [2,*], Mutian Yang [3]

[1] American International School of Budapest, Budapest, Hungary

[2] Tianyuan College, Hangzhou, China

[3] Sannew School, Wuhan, China

* Corresponding author email: 15020240212@xs.hnit.edu.cn

Abstract. Bias, which is one of the most significant social problems, can be reflected the Artificial intelligence. Those biases can be classified into three categories: Gender bias that shows the divergence between males and females, the religious bias that favors some beliefs over others, and racial bias that can be embodied in different AI's utilization efficiency by whites and nonwhites.

Keywords: AI; Bias; Gender; Race.

1. Introduction

Nowadays, technology is improving at an incredible rate, which assists our lives in many aspects. Artificial intelligence, in this case, comes into our views more frequently, which indeed bring lots of conveniences to our life. However, the world is full of bias. People are treated unequally due to some of their traits, such as race, gender, and religious beliefs. Unfortunately, those biases are also shown in artificial intelligence, which humans widely consider "just tools." Although most people have not realized the effects of these AI biases, many people are suffering from the discrepancy in using experiences. Without a doubt, these biases contained in artificial intelligence will severely influence our society and life as AI continues to develop rapidly in society. Under this situation, we conduct a detailed study about the effects of bias in artificial intelligence. Specifically, we divide biases in AI according to their traits: racial bias, gender bias, and religious bias. Throughout the study, we work on how those biases reflect in AI and how they will affect people's use of artificial intelligence.

2. Related Works

As mentioned previously, AI bias is a crippling obstacle in the way of full AI incorporation into society. However, although AI bias may seem inevitable due to the fact that the machine learning process inherently already contains bias, scientists and researchers have found potential solutions to this problem.

Firstly, a new method of AI learning- 'supervised learning' [1] has been proposed by scientists where the data learning pool fed to the AI is evaluated and filtered first by data scientists in order to ensure equal representation within the data, making the machine learning process mostly if not wholly bias-free.

On the other hand, instead of changing the method used for machine learning, some scientists have proposed a more straightforward solution by using 'synthetic data' [1]. Synthetic data sets are statistical representations of their current text counterparts. Since bias only exists within written text forms of media and cannot exist within pure statistics, replacing contemporary data forms with synthesized data might resolve the problem of AI bias. However, synthetic data is incorporated using statistics, so it struggles to represent more complex information properly. Thus, many scientists have deemed synthetic data to be an 'impractical' solution to AI bias. Furthermore, many scientists have resorted to a more long-term approach to solving AI bias. This is done by using 'explainable AI' [2] whenever explainable AI makes a decision based on its learning. It will tell the researchers why they made that decision.

So, if the decision was biased, the AI will elaborate on why it was biased during its decision-making process. Using this information, scientists could learn the main reasons why AI is biased in the long term. With this knowledge, scientists will eventually create a method to resolve AI bias.

In summary, many solutions to AI bias have been proposed. However, not all of them are applicable. Nevertheless, with each proposed solution and its methods, scientists are one step closer to solving AI bias. We are, slowly but surely, becoming more and more prepared for the bright, AI-assimilated future.

3. Method

In this paper, we take a comprehensive review of the effect of bias in AI community from two different perspectives: race and gender.

3.1 Race

The world is full of bias, and today's most-discussed bias among societies is the discrimination of race. Surprisingly, this discrepancy is also shown in artificial intelligence, which is widely considered "just tools." This can be traced back to the information sources of artificial intelligence, which are controlled and edited by high-level experts who consist of a higher proportion of Caucasians. These inputs of discriminatory information result in the different using effects of AI on people of different races. Here are some specific pieces of evidence to support my point.

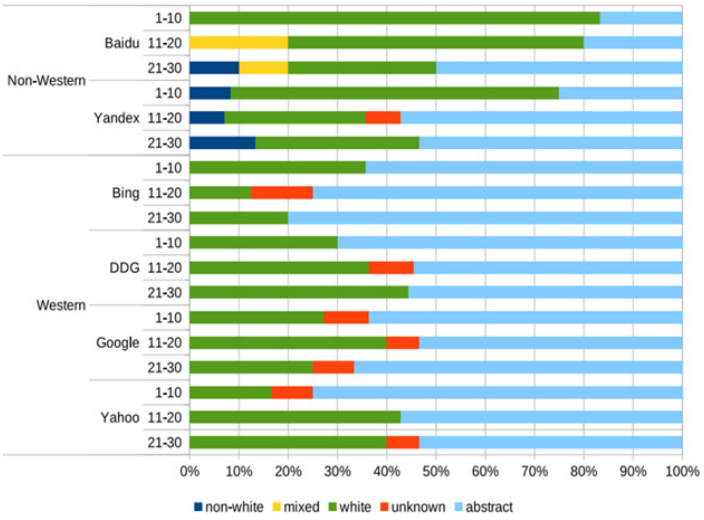


Figure 1. Proportion of racialized representations of AI.

Firstly, Caucasian figures in the image search outputs show some racial effects. When the keywords related to the human figure are typed in the search boxes, the pattern of results can be seen and summarized. In the study [3], Mykola, Aleksandra, and Roberto studied the pattern in search results and discovered some racial bias trends. In most of the search results, the human figure that came out first did not show any racial trait. Those are mostly abstract figures that are covered with blue skins and have appearances of technological styles. However, for some of it, pattern difference is discernible. Some of the search engines produced by the non-western company had shown a racial inclination. According to the research, those search engines showed mostly white subjects on top of the resulting web pages. For example, in the search engine named "Bing," the first ten results came out when searching for human figures composed of only abstract figures and those figures with white skins. Also, this phenomenon can even be seen more clearer in searches on "Baidu," which is the most famous search engine in China and is used by many Chinese. It resulted in the presence of a white figure. From the evidence above, a clear whiter pattern can be seen in these results. Also, this tendency can be identified from another aspect. Although the search results have not consisted only

of white figures and are combined with fictional figures, the fact that almost none of the search engines contain non-white results on not only the top of webpages but also results below can reinforce the point that artificial intelligence inclines white people. Thus, the image search result in the online search engine can indicate racial problems.

Secondly, the efficiencies in the usage of face recognition tools differ according to the users' races. In many fields, such as education, economy, entertainment, and social security, artificial intelligence's recognition of facial expressions plays an important role. This technology has become increasingly popular over time. Artificial intelligence finishes this task by learning through the deep learning method base on the dataset they gathered. However, with more and more collected data, some deviations are likely to appear in the procedure of acquiring data. Specifically, the accuracy in recognition of the faces in some races is worse than in other races' recognition results. This loss in precision seems to be due to insufficient human-figure data for those races. The data gathered by artificial intelligence have focused more on the groups of people with white skin. The research group experimented with clarifying the point to see if the dataset used by artificial intelligence influenced the result [4]. They repeatedly ask the artificial intelligence to learn datasets that combine with different races in different proportions through a deep learning process until they fully understand the data. After that, researchers use those AIs to see whether clear patterns are shown. Unsurprisingly, the race with minor data shows less accuracy in recognizing faces, and the races that are more familiar with AIs can be processed more effectively. Based on these data, a discrepancy effect caused by the unevenly distributed samples, which indeed exists in today's face reorganization tool, can be easily seen, reflecting the race problem.

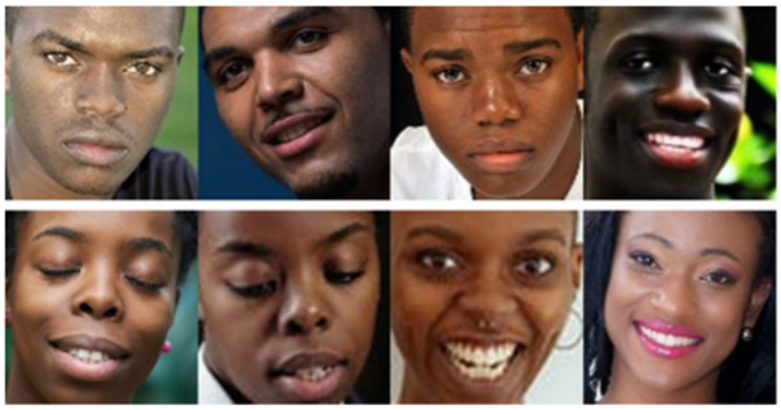


Figure 2. Images from Black dataset.



Figure 3. Images from East Asian dataset.

Thirdly, racial bias can also be seen in AI in the medical industry. As everyone knows, issue related to medical science has to be careful because it is related to a patient's health condition. Therefore, the artificial intelligence that helps in this field is always based on large amounts of data, and research

will cooperate with those data to reach their results. However, problems can have with the data. When scientists output data to artificial intelligence, they may unintentionally focus on specific groups of people, usually consisting of white people, and overlook others. In this case, the artificial intelligence that learned the incomplete data are very likely to misjudge the physical condition of patients belonging to the groups with insufficient data, which causes additional health problems and more medical expenses. For example [5], if an AI tests several black people's skin color and oximetry data of blood together, it may process this data with the standard for white people, which it gathered most in the stage of learning. In this way, black people, who have different physical conditions from white people, can be judged inaccurately by AI. When clinicians receive those data from artificial intelligence, they will wrongly estimate patients' health conditions and thus give them useless and even harmful prescriptions. Consequently, this may result in distrust between people and artificial intelligence and exacerbate race's social contradictions.

3.2 Gender

Although the negative connotations associated with prejudice, bias is just a criterion deviation. London and Danks [6] point out five biases in the automated system: Training data bias, Algorithmic focus bias, Algorithmic processing bias, Transfer context bias, and Interpretation bias.

Hence, AI bias is an abnormality in the results of machine learning algorithms caused by biases in the training data or discriminatory assumptions established during the algorithm development process. AI appears neutral, but it is caused by humans, which means it spontaneously internalizes all the same biases as modern people's Biases or discrimination - including gender bias.

The considerable evidence of sexual bias in the automated system is fought by us, exposing the gender disparities.

Regarding web search engines, Makhortykh [7] manifested that different web search engines can exhibit gender biases and a great difference in the Proportion of gendered representations of AI among the top 30 anthropomorphized images. To collect data, researchers utilized 200 virtual agents to record the result. Subsequently, they contain the top 30 images and are divided into five sets, male, female, mixed, abstract, and unknown. Then, they analyzed the statistics among five sets in distinct search engines. They observed that non-gendered portrayals are usually at a high ranking. Search engines tend not to attribute particular gender traits to it. In their top 10 results, most Western search engines found a preference for neutral pictures in gender, while gendered images show lower results. Yandex and BPS, by contrast, had more gender-specific images in the top 10 results. There was little male representation of AI that differs from Western engines. Both Baidu and Yandex emphasize males.

Lambrecht and Tucker [8] analyzed how advertisements promoting job opportunities in STEM fields are more often presented to men than women. This advertisement is neutral, apparently when the launch. However, the algorithm of modeling is to optimize the cost-benefit. Consequently, the ad shows more men because young women are essential to the target populations, which increases advertising costs.

In the emotional aspect, Domnich and Anbarjafari [9] analyzed gender bias. To assess the fairness of datasets, they used several model architectures as well as training and testing datasets. They evaluated feelings that were easier to distinguish between males and females. Simply, there are four facial expressions involved in their study (happy, sad, surprised, and upset) to ease the difficulty. There is another study about that. Cristina and Silvia [10] use only one model, but they include six basic principles of emotion (anger, fear, disgust, sadness, surprise, and happiness). These insightful studies suggest that using different algorithmic structures to identify different emotions leads to an inevitable preference for a particular gender.

4. Conclusion

This paper provided a comprehensive review of bias in artificial intelligence, which substantially expanded current knowledge of AI bias. Our study analyzed and combined three of the most dominant significant bias in AI that exerts an unavoidable impact on the development of AI algorithms.

Given that the AI bias is a bias that developed from the algorithm development process that may include the dataset bias and vague assumptions at the training, many potential biases influence the result. Consequently, extending this review work, the next objective has to be to expand research on other AI biases, which are a crucial factor that may arise a misrepresentation of AI dilation making and potentially exert adverse social impacts. Future research should focus on minimizing AI bias or designing a checking-bias algorithm. Potentially modifiable risk AI bias factors associated with the accuracy and discrimination of automated systems also need to be identified.

Hence, another direction of future work is to implement these solutions. After all, modern AI still involves a substantial amount of Uncertainty and hazards, so we must be prepared for the future.

References

[1] Apte, Poornima. "5 Ways to Prevent AI Bias." ITPro Today, 23 Sept. 2022, www. itprotoday. com/ artificial-intelligence/5-ways-prevent-ai-bias. Accessed23 Sept. 2022.

[2] Cannings, Nigel. "How To Tackle AI Bias." Spiceworks, 22 Sept. 2022, www. spiceworks. com/tech/ artificial-intelligence/guest-article/ how-to-tackle-ai-bias/. Accessed 23 Sept. 2022.

[3] Makhortykh, Mykola, Aleksandra Urman, and Roberto Ulloa. "Detecting race and gender bias in visual representation of AI on web search engines." International Workshop on Algorithmic Bias in Search and Recommendation. Springer, Cham, 2021.

[4] Sham, Abdallah Hussein, et al. "Ethical AI in facial expression analysis: Racial bias." Signal, Image and Video Processing (2022): 1-8.

[5] Leslie, David, et al. "Does “AI” stand for augmenting inequality in the era of covid-19 healthcare?." bmj 372 (2021).

[6] Danks, D. and London, A.J. 2017. Algorithmic Bias in Autonomous Systems A Taxonomy of Algorithmic Bias. 26th International Joint Conference on Artificial Intelligence (IJCAI-17). Ijcai (2017), 1–7. DOI: https: //doi.org/10.24963/ijcai.2017/654.

[7] Makhortykh, M., Urman, A., Ulloa, R. (2021). Detecting Race and Gender Bias in Visual Representation of AI on Web Search Engines. In: Boratto, L., Faralli, S., Marras, M., Stilo, G. (eds) Advances in Bias and Fairness in Information Retrieval. BIAS 2021. Communications in Computer and Information Science, vol 1418. Springer, Cham. https://doi.org/10.1007/978-3-030-78818-6_5.

[8] Lambrecht, A. and Tucker, C. 2019. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. Management Science. 65, 7 (2019), 2966–2981. DOI: https: // doi.org/10.1287/mnsc.2018.3093.

[9] Domnich, A. and Anbarjafari, G. 2021. Responsible AI: Gender bias assessment in emotion recognition. (2021), 1–19.

[10] Ekman, P. 1992. An argument for basic emotions. Cognition and Emotion. 6, 3–4 (1992), 169–200. DOI: https:// doi.org/10.1080/02699939208411068.