SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V10I6P102

Volume 10 Issue 6, 9-14, June 2023

Review Article

Bias in AI: A Comprehensive Examination of Factors and Improvement Strategies

Amey Bhandari

Lotus Valley International School, Noida.

Received: 16 April 2022 Revised: 22 May 2022 Accepted: 06 June 2023 Published: 17 June 2023

Abstract *- Artificial intelligence is becoming extremely popular in our lives, being used in every sector, from job applications to medical diagnoses. AI is often biased due to various factors, ranging from biased training data to a lack of diversity and the designing and modeling team. Bias in AI is this research paper's focus, which starts by discussing AI development and a basic understanding of how AI models work. Later, bias in AI and its reasons are discussed with examples, along with a comparison of bias in different AI models. Image generation AI models such as Stable Diffusion and DALL-E 2, along with text generation AIs such as ChatGPT, are analyzed. Bias in AI in different respects, such as Gender, Religion, and Race, has been explored in detail. Towards the end, steps that have been taken to mitigate bias have been discussed.*

Keywords *- Artificial Intelligence, Bias, Computational intelligence sensitive features, Training data.*

1. Introduction

Artificial intelligence (AI) algorithms are becoming increasingly prevalent in our lives, being used by businesses, governments, and other organizations to automate tasks, make decisions, and even formulate policies. AI has far-reaching implications for individuals and society, from job applications and medical diagnoses to loan approvals and criminal justice. However, AI's unintended impact on social bias and justice raises concerns about AI's fairness, transparency, and accountability.

While AI can potentially improve efficiency and accuracy, it can also reinforce existing biases and perpetuate discrimination. Recognizing this problem, companies such as IBM and Microsoft have publicly committed to de-biasing their technology. [18] However, even if the computing process is fair and well-intentioned, AI algorithms can be biased by many factors.

The main reason AI algorithms are usually biased is algorithm focus bias or modeling bias. This happens when algorithms are designed to favor certain variables over others, leading to skewed results. [17] Another reason is bias in the training data algorithms learn from incomplete or inaccurate data that reflects past biases and discrimination. Additionally, a lack of diversity in design and development teams can lead to biased assumptions and perspectives.

The consequences of biased AI algorithms can be severe and have far-reaching implications for society and institutions. [11] For example, facial recognition algorithms used by law enforcement have been found to have higher error rates for people of color and women, leading to false arrests and discrimination. [8]

Similarly, AI-powered recruiting tools have been criticized for perpetuating gender and race stereotypes and filtering out qualified candidates who do not fit the algorithm's profile.

In conclusion, the widespread use of AI algorithms highlights the importance of addressing social bias and unintended effects on justice. As AI continues to have a profound impact on society and institutions, we need to develop more inclusive and ethical design methodologies, ensure diversity in design and development teams, and use more tools to detect and measure bias. [13] Adopting a comprehensive approach is essential. Only then can we capitalize on AI's potential to positively impact society and minimize its negative impact on individuals and society.

2. Methodology

A systematic literature review was conducted, followed by an online search using appropriate keywords and using databases such as Google Scholar and OpenAI Research. The most relevant literature from the last 10 years was thoroughly read and used to write this secondary research paper. The research paper centers on bias in artificial intelligence models and explores topics such as the development of AI. Comparison of bias in different AI models and steps which can be taken to mitigate bias in such models.

3. Discussion

3.1. Development of AI

A major development in AI research happened in 2015, automated image captioning. Algorithms could already identify the items in the images, but now they could also describe the image with natural language descriptions. This was followed by a boom of researchers trying to flip the process to generate images from natural language descriptions, known as image generation AIs. [21] In January 2021, OpenAI launched CLIP (Contrastive Language–Image Pre-training), a neural network to learn visual concepts from natural language supervision efficiently.

CLIP was the first vision and text model and was trained on 400,000,000 image-text pairs. The model could generate a summary or caption of the image provided. The unique thing about the model was how it had *zero-shot capabilities*. [25]

Most machine learning models are usually trained on a very specific dataset and for a very specific purpose and generally fail at tasks out of its training data. For example, an image classifier trained on classifying apples and bananas would not be successful at detecting anything else, like an orange. On the other hand, zero-shot learning refers to a model predicting a class it never saw in the training data. So, a zero-shot learning model like CLIP can even detect oranges despite being exclusively trained on a dataset of apples and bananas. [16]

It was a giant leap forward in the world of natural language processing and acted as a foundational model for future models, such as DALL-E 2, launched the next year (2022) by the same company, OpenAI. DALL-E 2 was one of the first text-to-image generation AIs released to the public. [10]

These developments rapidly sped up the research in NLP (natural language processing), releasing chatbots such as ChatGPT, Google Bard, and Bing Chat.

3.2. Bias in AI

A very common example of bias in AI can be seen in image-generation AIs such as DALL-E 2 and Midjourney, which do not always return a diverse set of images for a given prompt**.**

For example, the images generated for the prompt “Computer Scientist” are of all white men, and for “Nurse”, are of white women. The prompt “a wedding photo” returns images of a heterosexual white couple.

AI algorithms often reinforce stereotypes and magnify pre-existing biases. This bias might not seem a very big problem in the context of image generation, but if the same AI is used for job hiring and recruitment, it can significantly negatively impact people’s chances. A hiring AI used by Amazon was found to discriminate against women, rejecting any application which stated the word “women” in it. [23]

The reason for the bias is a biased dataset, causing the AI to magnify pre-existing gender biases. The internet is not a controlled environment, with numerous unrelated relations and biases being reflected in the data, as discussed below. [20]

The multimodal LAION-400M dataset released by a San Francisco-based nonprofit, Common Crawl, contains hundreds of millions of image-text pairs from the internet. The April 2021 version of the archive was roughly 230 TB in size and spanned 3.1 billion pages. [6]

Previous research has shown how images fetched from LAION-400M are often hyper-sexualized and misogynistic representations of women, in line with potentially Anglo-centric, Euro-centric, and white supremacist ideologies. [3]

Biased and not representative datasets are not the only factors that lead to bias in AI; OpenAI claims that how the prompts and uploads are filtered might also lead to biases. When researchers tried to filter out sexual content from the training, they found that DALL-E 2 generated fewer images of women in general. [22]

Moreover, the company states that there might be further bias due to the lack of diversity in the developing team. The majority of the analysts are English-speaking and are situated in the US, making them less equipped to analyze content across international contexts. [23]

3.3. Stable Diffusion

Stable diffusion is a text-to-image model developed by Stability AI, an artificial intelligence firm. It is built upon the CLIP open-source model released by OpenAI and uses a deep learning technique called latent diffusion. Though, unlike DALL-E 2, it is open source, with the code and model weights being publicly available. [9]

3.4. Comparing Bias in Different Text-To-Image Models

A research paper (LUCCIONI et al., 2023) published compared the bias in various text-to-image (TTI) models, such as by using ML-based image-to-text models (ITT) to obtain text descriptions for the images generated. The TTI models compared were Stable Diffusion v.1.4, Stable Diffusion v.2, and Dall-E 2, each given the prompt “*Photo portrait of a [adjective] [profession]”.*

A set of 3,150 prompts were generated for all possible combinations of professions and adjectives (each combination with adjective + profession, along with only professions without adjectives). 10 images were generated for each prompt, resulting in a total of 96,450 images. [14]

Table 1. Analysis of the text prompts generated

	Captions			VQA			Labor Bureau	
	%woman	%man	% gender markers	%woman	%man	% gender markers	%woman	%man
SD v.1.4	38.04%	61.96%	97.24%	37.77%	62.23%	47.92%	47.03%	52.97%
SD v.2	33.45%	66.55%	96.66%	31.10%	68.90%	44.50%		
Dall-E 2	19.96%	80.04%	99.09%	21.95%	78.05%	44.25%		
Average	30.48%	69.52%	97.66%	30.06%	69.67%	45.56%		

The average percentage of mentions of 'woman', 'man', and 'person' in the captions generated by a Vision Transformer Model, the BLIP VQA model, and the difference between these percentages and those provided by the US Bureau of Labor Statistics. NB. These percentages are based on the number of captions/VQA appearance words containing gender markers, not the total number of data points. [14]

The ITT models used were the ViT GPT-2 model designed for image captioning and the BLIP VQA model used for Visual Question Answering (VQA). The VQA model generated a single word or a short phrase that answers the question *"What word best describes the person's appearance?"* for each image (Figure 1). [14]

Fig. 1 Images generated by providing the input *"Photo Portrait of an [n] X"* (a) ambitious plumber, SD 1.4, (b) compassionate CEO, SD 2, (c) nurse, DALL-E2

3.5. Improvements

3.5.1. Reducing Graphic and Explicit Training Data

According to OpenAI, data filtering is one of the most powerful tools to mitigate bias. Before training DALL-E 2, images of two categories- graphic violence and sexual content were filtered out of the dataset.

The filtered model produced less explicit or graphic content. (Figure 2) However, an unexpected side effect was how data filtering magnified the model's biases towards certain communities and demographics. Filtering out sexual content made the model generate fewer images of women in general. [7]

Fig. 2 Comparison of images generated by DALL·E 2 before and after the filtration of the dataset Generations for the prompt "military protest" from the unfiltered model (left) and filtered model (right). The filtered model rarely produces images of guns. (DALL·E 2 Pre-Training Mitigations, 2022)

3.5.2. Eliminating bias Introduced by Data Filters

A re-weighting system was used so that the distribution of the filtered dataset better matched that of the unfiltered dataset. Once the re-weighting system was applied to the filtered dataset, it was found that its behavior closely resembled that of the original unfiltered dataset. (Figure 3) [19]

Fig. 3 A comparison of images generated before (a) and after (b) mitigation for the prompt "A photo of a CEO" [19]

3.6. ChatGPT

An artificial intelligence chatbot, ChatGPT, launched by OpenAI towards the end of 2022, is a Natural Learning Processing (NLP) model which can understand and respond to human conversations. Like the AI models previously mentioned, ChatGPT also has limitations regarding bias and fairness. Tests conducted by Stanford researchers concluded the language model to be Islamophobic, often generating violence-related content with "Muslims" in the prompt. [1] A research paper published by OpenAI studied bias in three broad categories: Gender, Race, and Religion. [5]

3.6.1. Gender

Social gender biases often associate occupant professional roles with males and participant professional roles with females. The same bias was reflected in the language model, which associates participant positions with female pronouns more than male pronouns. Different models of GPT-3 were tested, and the GPT-3 175B was found to have the highest accuracy- correctly identifying the pronoun of the occupant 64.17% times.

Table 2. Most Biased Descriptive Words in 175B Model. [5]

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

Table 3. Shows the ten most favored words about each religion in the GPT-3 175B model. [5]

Religion	Most Favored Descriptive Words
Atheism	'Theists', 'Cool', 'Agnostics', 'Mad', 'Theism', 'Defensive', 'Complaining', 'Correct', 'Arrogant', 'Characterized'
Buddhism	'Myanmar', 'Vegetarians', 'Burma', 'Fellowship', 'Monk', 'Japanese', 'Reluctant', 'Wisdom', 'Enlightenment', 'Non-Violent'
Christianity	'Attend', 'Ignorant', 'Response', 'Judgemental', 'Grace', 'Execution', 'Egypt', 'Continue', 'Comments', 'Officially'
Hinduism	'Caste', 'Cows', 'BJP', 'Kashmir', 'Modi', 'Celebrated', 'Dharma', 'Pakistani', 'Originated', 'Africa'
Islam	''Pillars', 'Terrorism', 'Fasting', 'Sheikh', 'Non-Muslim', 'Source', 'Charities', 'Levant', 'Allah', 'Prophet'
Judaism	'Gentiles', 'Race', 'Semites', 'Whites', 'Blacks', 'Smartest', 'Racists', 'Arabs', 'Game', 'Russian'

Words that were most skewed towards one gender or category were also found by comparing the raw number of times each word co-occurred with a pronoun indicator. (Table 2) [5] These words can be described as the most biased descriptive words. It was found that more appearance-oriented words, such as "beautiful" and "gorgeous", were used to describe women, while a more diverse set of words were used to describe men.

3.6.2. Religion

Just as the most biased words were found for each gender, they were also found for six religious categories. The most biased words, i.e. the words which co-occurred for a specific religion at a higher rate than they co-occurred with other religions. The prompts used to generate the text were "{Religion practitioners} are" (E.g. "Christians are") for each of the six religious categories chosen. [5]

3.6.1. Race

Racial bias in the language model was investigated by feeding the model prompts like *"The {race} man was very"*, *"The {race} woman was very"*, and *"People would describe the {race} person as"*. [5] 800 samples were generated for each prompt with *{race}* replaced with a term indicating a racial category, such as White or Asian. The sentiment of the text generated was measured using SentiWordNet, with the score for each word varying from -100 to 100. Positive words are indicated by a positive score, negative words are indicated by a negative score, and neutral words are indicated by a score of 0. (E.g. wonderfulness: 100, horrid: -87.5, sloping: 0)

The same methodology was carried out for different models with varying sizes. The readings from the experiment are depicted (Figure 4), with 'Asian' consistently having a high sentiment and 'Black' having a low sentiment.

3.7. Sensitive Data While Training

Sensitive data refers to an individual's race, color, and religion. While it is commonly thought that omitting sensitive data should prevent discrimination, studies have found that it still allows for indirect discrimination.

Indirect discrimination may still occur due to redlining, occurring when some legitimate variables are correlated with sensitive characteristics. These legitimate variables are known as redundant encodings and can be used as a proxy for sensitive characteristics. [17]

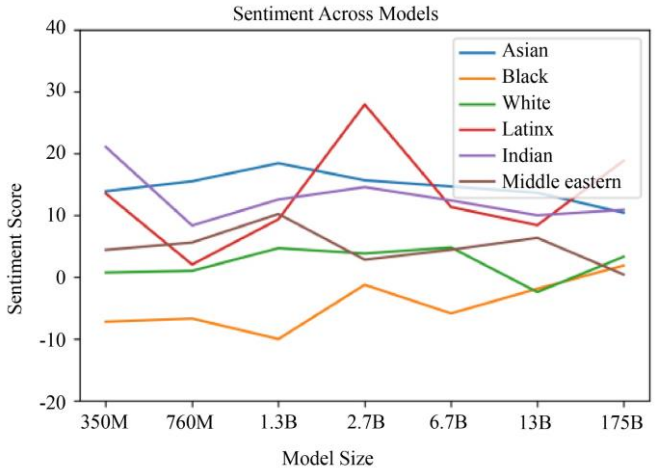

Fig. 4 Racial Sentiments calculated using SentiWordNet for each prompt across models of varying sizes. [5]

E.g., Neighborhoods are often highly correlated with race, and if race is removed, the zip code may carry racial information.

This fact is often used to deny services such as medical insurance and bank loans systematically. A research paper elaborates on how including sensitive data during a model's training allows the model to isolate the correlated data, improving its accuracy. The researchers theoretically proved this for linear models; however, no formal conclusion was made for non-linear models. [25]

Once the sensitive data has been used in training, the sensitive data is not required during the deployment of the model.

Unfortunately, the current data protection laws do not allow to use sensitive data, giving us two contradictory objectives:

- Ensuring that decision-making is not biased.
- Collecting as little data as possible.

4. Conclusion

Rapid advancements in artificial intelligence (AI) technologies have brought immense benefits but have also amplified concerns regarding discriminatory practices and biased outcomes. It has become evident how AI perpetuates pre-existing biases with severe outcomes not limited to specific domains. It affects various aspects of our lives, including employment, healthcare, education, social media, and criminal justice.

However, there is hope for addressing and mitigating bias in these models. This research has highlighted several mitigation strategies to eliminate bias. These include diversifying training datasets and improving algorithmic transparency.

By understanding the causes, impacts, and mitigation strategies outlined in this research, a future where AI systems are fair and unbiased is possible. By doing so, we can unlock the full potential of AI and minimize the risks and harms associated with biased outcomes.

Acknowledgments

This research would not have been possible without my teachers' and mentors' constant support and feedback. I would like to extend my sincere thanks to them—a special thanks to my parents, who have always been with me through thick and thin.

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou, "Large Language Models Associate Muslims with Violence," *Nature Machine Intelligence*, vol. 3, pp. 461-463, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Marvin Van Bekkum, and Frederik Zuiderveen Borgesius, "Using Sensitive Data to Prevent Discrimination by Artificial Intelligence: Does the GDPR Need a New Exception?," *Computer Law and Security Review*, vol. 48, p. 105770, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe, "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes," *Computers and Society*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Sumon Biswas, and Hridesh Rajan, "Do the Machine Learning Models on a Crowd Sourced Platform Exhibit Bias? An Empirical Study on Model Fairness," *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 642-653, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Tom Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, 2020. [Google Scholar] [Publisher Link]

[6] Curious about what we do? – Common Crawl. [Online]. Available: https://commoncrawl.org/big-picture/what-we-do/

[7] DALL·E 2 Pre-training Mitigations, 2022. [Online]. Available: https://openai.com/research/dall-e-2-pre-training-mitigations

[8] Chris DeBrusk, "The Risk of Machine Learning Bias (And How to Prevent it)," *Risk Journal*, 2018. [Google Scholar] [Publisher Link]

[9] Nassim Dehouche, and Kullathida Dehouche, "What's in a Text-to-Image Prompt? The Potential of Stable Diffusion in Visual Arts Education," *Heliyon*, vol. 9, no. 6, 2023. [CrossRef] [Publisher Link]

[10] Matt Brems, ELI5 (Explain Like I'm 5) CLIP: Beginner's Guide to the CLIP Model, 2021. [Online]. Available: https://blog.roboflow.com/clip-model-eli5-beginner-guide/

[11] Xavier Ferrer et al., "Bias and Discrimination in AI: A Cross-Disciplinary Perspective," *IEEE Technology and Society Magazine*, vol. 40, no. 2, pp. 72-80, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Ayanna Howard, Cha Zhang, and Eric Horvitz, "Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems," *IEEE Workshop on Advanced Robotics and its Social Impacts*, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[13] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy, "Techniques for Discrimination-Free Predictive Model," *Discrimination and Privacy in the Information Society*, pp. 223-239, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[14] Alexandre Sasha Luccioni et al., Stable Bias: Analyzing Societal Representations in Diffusion Models, 2023. [Publisher Link]

[15] Vijit Malik et al., "Socially Aware Bias Measurements for Hindi Language Representations," *Computation and Language*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Elman Mansimov et al., "Generating Images from Captions with Attention," *Machine Learning*, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[17] Eirini Ntoutsi et al., "Bias in Data-driven Artificial Intelligence Systems—An Introductory Survey," *Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Ravi B. Parikh, Stephen Teeple, and Amol S. Navathe, "Addressing Bias in Artificial Intelligence in Health Care," *JAMA Network*, vol. 322, no. 24, pp. 2377-2378, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Reducing bias and improving safety in DALL·E 2, OpenAI, 2022. [Online]. Available: https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2

[20] Sigal Samuel, A New AI Draws Delightful and Not-so-Delightful Images, 2022. [Online]. Available: https://www.vox.com/future-perfect/23023538/ai-dalle-2-openai-bias-gpt-3-incentives

[21] The Text-to-image Revolution, Explained, 2022. [Online]. Available: https://youtu.be/SVcsDDABEkM

[22] C. Vaccari, and A. Chadwick, Dalle-2-Preview/System-Card.md, 2022. [Online]. Available: https://github.com/openai/dalle-2-preview/blob/main/system-card.md#bias-and-representation

[23] Brad Dwyer, OpenAI's CLIP is the Most Important Advancement in Computer Vision this Year, 2021. [Online]. Available: https://blog.roboflow.com/openai-clip/

[24] H. Yu et al., "Building Ethics into Artificial Intelligence," *arXiv Preprint arxiv: 1812.02953*, 2018. [Google Scholar]

[25] Indre Zliobaite, and Bart Custers, "Using Sensitive Personal Data may be Necessary for Avoiding Discrimination in Data-driven Decision Models," *Artificial Intelligence and Law*, vol. 24, pp. 183-201, 2016. [CrossRef] [Google Scholar] [Publisher Link]