A survey on Artificial Intelligence Bias

Darsh Damani

Received June 21st, 2022
Accepted October 11th, 2022
Electronic access November 30th, 2022

Artificial Intelligence(AI) is beneficial to society as it can reduce the time to perform tasks.Artificial intelligence bias refers to results that are systemically prejudiced due to erroneous assumptions in the machine learning process;it can enter through different sources and reduce its benefits. A great deal of bias stems from human, systemic, institutional biases. AI bias may prevent widespread adoption of AI in society, as it may then affect certain sections of society negatively. This manuscript is about how bias can enter a model as well as provides methods on how to mitigate it.

Bias can be classified in three ways: pre-existing, emergent, and technical. Bias can enter the model in two ways, which are issues with representing the goal of the model or issues with the dataset. Ways to mitigate bias have been described in detail and have been split into three main categories which are post-processing, pre-processing, and in-processing. Relevant data from leading research papers has been organized in an understandable manner suitable to those who are not yet familiar with the topic but aspire to be. It goes in-depth on each of the topics mentioned as well as provides clear and relevant information on each topic. Relevant figures for understandability have been provided throughout the paper.

To summarize, this manuscript has brought attention to what AI bias is, where it comes from, and how to mitigate it. This manuscript is about the recent developments and advancements in Artificial Intelligence, specifically about how it can enter a model and ways to mitigate it.

Introduction

Bias is the prejudice against one person or group, especially in a way considered to be unfair. Artificial Intelligence bias refers to the assumption made by the model and reflects the author's choice of data algorithm, blending methods, model construction practices and how it is applied and interpreted[1].

Due to bias, we cannot adopt some models into real-world applications. For example, Amazon's recruiting engine which turned out to be biased against women in the process.Amazons computer models were trained to vet applicants by observing patterns in resumes submitted to the company for over ten years. Most came from men, a reflection of male dominance across the tech industry. It automatically handicapped the resumes that contained words like "women" and also automatically downgraded the graduates of two all-women colleges. Due to this, amazon decided to not use the model[2]. Minimizing bias in AI will be crucial in increasing our trust in it, which in turn will allow AI to be used more in real world applications.

We present a literature review summarizing some of the ways bias can creep inside the paper as well as different ways to combat it. We have mentioned different techniques some authors have developed as well as mentioned what techniques are used in AI360F, a toolset developed by IBM. This paper fits in as a guide for those who are not yet familiar with bias and want to get knowledge on some breakthroughs in it.

This paper is organized into three sections. Section I talks about what is bias and why it is an issue. Section II talks about how bias can creep inside a model. Section III talks about ways to mitigate bias.

Classifying bias

Artificial intelligence bias is a reflection of the data algorithm the authors choose to use as well as their data blending methods, model construction practices, and how the model is applied and interpreted. It refers to the assumptions made by a model[4].

Bias can be classified into three types: pre-existing, technical, and emergent. It is classified based on the stage in which bias enters the system.

Pre-existing bias

It is when the system embodies biases that exist before the creation of the system. They may reflect the biases of those who have significant input in the creation of the system, for example, the system designer or client. It can enter a system either through the conscious efforts of individuals or institutions or unconsciously and implicitly[3]. For example, a system that advises on loan applications, the system negatively weighs applicants who live in an "undesirable" location. The

Table 1 Main differences between Pre-existing bias, Technical bias and Emergent Bias.

	Pre-existing bias	Technical bias	Emergent bias
Where they come from.	Biases that are present before the creation of the model are called pre existing biases.	Technical biases are found in the design process and include limitations of computer tools such as hardware and software, peripherals, and imperfections in number generation.	Biases that arise in a context of use,
Why does it occur?	They reflect the biases of those who had significant input in the creation of the system.	It occurs when we quantify the qualitative, discretize the continuous, or formalize the informal[3].	It emerges as a change in societal knowledge, population and culture and cannot be identified from before.

program embeds biases of clients or designers who want to avoid certain applicants due to stereotypes[3].

Technical bias

It is found in the design process and includes limitations of computer tools such as hardware and software, peripherals, and imperfections in number generation. It occurs when we quantify the qualitative, discretize the continuous, or formalize the informal[3]. An example is a technical constraint imposed by the size of the monitor screen forces the presentation of flight options, thus, making the algorithm chosen to rank flight options critically important. Whatever ranking algorithm is used, if it systematically places certain airlines' flights on initial screens and other airlines' flights on later screens, the system will exhibit technical bias[3].

Emergent bias

It arises in a context of use, it emerges as a change in societal knowledge, population and culture and cannot be identified from before[3]. An example is an automated airline reservation system that envisions a system designed for a group of airlines that serve national routes. If the system was extended for international airlines, it would place the airlines at a disadvantage. User interfaces are likely to be particularly prone to emergent bias because interfaces by design seek to reflect the capacities, character, and habits of prospective users[3].

Ways bias can creep inside

Bias can further enter the model in many different ways. They are mainly divided into issues representing the goal or datasets.

Issues with representing the goal

The earliest stage bias can creep in is when the deployer is deciding the objective of the model. It is further divided into proxy goals, feature selection, and limitations due to hardware[5]. Proxy goals are basing the model on the goal and data on historic information without factoring in any biases involved[5]. An example of the above is in advertisements, there is no way to determine the likelihood of a product being purchased. They may choose attributes similar to that of previous customers. It is not a correct approach as the data may have a historic bias in it. Companies must choose input attributes, and training labels, and hypothesize and reinforce criteria that they feel are the best[5].

Feature selection occurs due to the choice of which attributes to include and not include may lead to bias. Harmless features may impact the model positively or negatively which tend to remain hidden. Features which are not included but may favorably influence the predictions for some people are even more difficult to quantify[5].

Limitations due to hardware occurs as artificial intelligence requires a lot of math and processing power. Training sets are numerically represented and are large and have features obtained on a large scale. Mathematical reductions result in information loss. It may serve as a proxy for restricting data[3].

An example of the above is a job screening service that may use credit scores favorably and if a person does not insert it, it may look on it negatively. It may not look at the information that is harder to quantify such as letters of recommendation[3].

Issues with training data and production data

It may create problems during the mapping phase. Creating datasets is time-consuming and usually involves using a large dataset and to perform supervised learning, getting training labels is time-consuming[6]. Due to the scale and complexity, creating a dataset is often the source of many problems[7]. This can manifest itself in different ways such as unseen cases, mismatched data sets, irrelevant correlations, and non-generalizable features.

Unseen cases occur due to Artificial Intelligences's ability to generalize solutions robustly. This becomes a problem when the model doesn't know what to do. Algorithms can result in bias when they are used in a situation they are not intended for. It becomes a disadvantage for the group(class) which hasn't been trained and can lead to hidden mispredictions which impact a group negatively[5]. An example is if a model is trained in English, but is shown German, it won't know what to do. The potential misinterpretation of an algorithm's outputs can lead to biased actions through what is called interpretation bias. For example, algorithms utilized to predict a particular outcome in a given population can lead to inaccurate results when applied to a different population[5].

Mismatched data sets occur when the training data may not match data from real-world applications; this occurs when the given data is different from the testing data. There are possibilities that training data may change over time which can have hidden effects on the model[5]. An example is a commercial facial recognition system trained on mostly fair-skinned subjects that have vastly different error percentages for different populations: 0.8% for lighter-skinned men and 34.7% for darker-skinned women[8].

Irrelevant correlations occur due to training data having correlations between irrelevant features. The distribution of irrelevant correlations may not be particular to the training set but may occur in real-world data[5]. For example, Ribeiro et al. trained a classifier to differentiate between dogs and wolves with images of wolves surrounded by snow and dogs without snow. After training, the model sometimes predicts that a dog surrounded by snow is a wolf[9]. Unlike non-generalizable features, these may be there in the training data as well.

Non-generalizable features means that the training data is highly curated and real-world data is often corrupted or incomplete and very rarely curated. Stale data used for training and production input may be outdated[5]. For example, credit scores could be downloaded from an external source and stored locally for fast access. Unfortunately, there may be resistance to updating the dataset by developers as it may reset the baseline for ongoing training experiments.

Ways to mitigate bias

Lots of research goes into developing techniques to mitigate bias. There are three possible places for intervention which are preprocessing, in processing and postprocessing.

Preprocessing methods

Preprocessing methods are approaches that mainly focus on the data and try to produce a balanced dataset. The fairer the dataset, the less biased the model will be, resulting in the least prejudice and unfairness.The more fair a dataset gets, issues such as mismatched datasets, unseen cases, feature selection all reduce and are less prone to be the reason why the model is biased, if it is after using a fair dataset. Designers not only examine the design specifications but must couple this examination with a good understanding of relevant biases out in the world. Thinking about bias should be there in the earliest stages, such as negotiating the system requirement with the client. The computing community is developing and understanding bias mitigation techniques, and we can correspondingly develop or apply these techniques to minimize bias. Decisions here include how to frame the problem, the purpose of the AI component, and the general notion that there is a problem requiring or benefitting from a technology solution. They include substantiation, and vetting of the training data.

Substantiation refers to providing evidence for the hypothesis.For example ,simulated data were generated in order to substantiate a hypothesis, and the results obtained from the analysis seemed to support it. It requires preparation to provide quantitative evidence for the validity of your chosen numerical representations, the hypothesis itself, and the impact of the application on its environment, including its future input. When surrogate data is used, it should be accompanied by quantitative evidence that suggests that the surrogate data is appropriate for its intended use in the model. Limitations of the surrogate data should be documented and presented during reviews of predictions

Vetting the training data refers to examining the dataset for accuracy, relevance, and bias. Incomplete or vague samples need to be removed. The time and effort required to make the dataset of good quality may not be worth it asproduction data may change over time and there are some samples that will be ambiguous and vague. Curated datasets may not work with production data and data may have been manipulated while vetting.

Kamiran and Calders have proposed methods which introduce a new classification scheme for learning unbiased models on biased training data. They propose the least intrusive

Table 2 Summarizing and differentiating between Pre-processing, In-processing and Post-processing methods.

	Pre-processing methods	In-processing methods	Post-processing methods
What do these methods mainly focus on?	They are methods that mainly focus on the data and trying to produce a fair dataset.	These methods focus more on the model and they tackle the classification problem by integrating the model's discriminative behavior in the objective function through regularization or constraints, or by training on target labels[10].	They are methods that correct system errors in model output by comparing hindcasts to observations. They generally take a subset of samples and change their predicted labels appropriately to meet a group fairness requirement.
Examples of the methods included in the paper.	Examples of some methods include substantiation, and vetting of the training data.	Examples of some methods include detecting data divergence, establishing processes to test for bias, preventing technical bias, optimization over context, and adversarial debaser.	Examples of some methods are matching the production data to training data, preventing emergent bias, and humans in the loop.

modifications which lead to an unbiased dataset. On the modified dataset they use a non-discriminating classifier[11]. For the results, the methods are able to reduce the prejudicial behavior for future classification significantly without losing too much predictive accuracy[12]. Kamiran and Calders have also worked with input data containing unjustified dependencies between some data attributes and the class label and solved the problem by finding an accurate model for which the predictions are independent from a given binary attribute[13] or by carefully sampling from each group[12]. Calmon, Wei, Vinzamuri, Ramamurthy, and Varshney[14] proposed a probabilistic fairness-aware framework that alters the data distribution towards fairness while controlling the per-instance distortion and preserving data utility for learning.

AI 360F which is a toolset developed by IBM uses several methods- reweighing, optimized pre-processing, learning fair representations and disparate impact remover. Re Weighing generates weights for each training example, group or label, differently to ensure fairness. Optimized preprocessing[15] learns a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives. Learning fair representations[16] finds a latent representation that encodes the data well but obfuscates information about protected attributes. Latent representation is a machine learning technique that attempts to infer variables that cannot be inferred directly(latent variables) through empirical measurements.

Disparate impact remover[17] edits feature values to increase group fairness while preserving rank-ordering within groups.

1 In Processing Methods

In-processing methods are approaches that tackle the classification problem by integrating the model's discriminative behavior in the objective function through regularization or constraints, or by training on target labels[10]. Processing methods such as detecting data divergence, establishing processes to test for bias, preventing technical bias, optimization over context, and adversarial debarring.

Detecting data divergence refers to the technique of actively monitoring for incomplete data (especially when the model is trained on clean data)[5]. For example, an employer recruiting program may use credit scores to train a filter to screen candidates and confuse someone with no credit history with someone with a low credit history[5].

Establishing processes and practices to test for and mitigate bias in AI systems refers to operational procedures that can include improving data collection through more knowledgeable sampling and using third parties to inspect data and models, as well as proactively engaging with communities affected. Transparency about processes and metrics can help the community understand the steps taken to promote fairness and any associated trade-offs. Teams are normally equipped with frameworks that allow them to prioritize equity when defining their objects.Ensure that datasets are used responsibly and labeled and ensure variables do not disadvantage anyone. These ensure responsible algorithmic development[18].

To prevent technical bias, a designer must envision the design, the algorithms, and the interfaces in use so that decisions do not run at odds with moral values[3]. For example, even the largely straightforward problem of whether to display a list with random entries or sorted alphabetically,, a designer might need to weigh considerations of ease of access enhanced by a sorted list against constraints afforded by the hardware such as processing power used[3].

Optimization over context occurs when designers are focused on the system's accuracy and performance which may result in bias in the model.The ecological fallacy occurs when an inference is made about an individual based on their membership within a group. Unintentional constraints can cause results that reinforce societal inequities. These inequities help in increasing the model's accuracy, hence enabling the research community to discover them would be a way to manage them[19]. They serve as a positive effect of algorithmic modeling. For example-university admissions algorithm GRADE, which was shown to produce biased enrollment decisions for incoming Ph.D. students Without ground truth for what constitutes a "good fit," a construct, was developed using prior admission data. Once put into production, the model ended up being trained to do a different job than intended. Instead of assessing student quality, the model learned previous admissions officer decisions[19] which may have had biases in them and could be partial to certain groups. Another issue is that candidate quality cannot be truly known until after the student matriculates.

2 Post-processing

Preventing emergent bias means the designers should know the context of use and design accordingly. It is important to anticipate domains prone to bias, such as previous examples of biased systems and data. This is very important in applications that are likely to be adopted. When it is not possible to design for extended use, the designers should attempt to articulate constraints for the appropriate use of a system. They should take action and be responsible if bias emerges.There is a high risk that AI can exacerbate the bias[20].

Diverse and multidisciplinary teams which include but aren't limited to men, women, and minorities should be included to work on the data.Humans in the loop involves engaging individuals in the social sciences and humanities – as well as domain experts that understand the particular domain the AI system is meant to operate in.A few practices involving humans in the loop include- measuring the sharing the data on diversity, investing, involving domain experts and exploring how machines and man work together and can reduce bias. . Another suggestion is the machine may provide recommendations for the humans involved which keeps the humans involved while taking the machine's help. Transparency about the algorithm may help in how much weightage the humans assign to the recommendations provided by the model. It is important to enable a culture that prioritizes equality over accuracy when it is not feasible to mention the shortcomings. Performance reviews should include a component around ethical practices. Embed training on ethics, bias, and fairness for employees developing, managing, and/or using AI systems[18,21].

Whitebox methods post-process the classification model once it has been learned from data. This consists of altering the model's internals. Examples of the white-box approaches consist of correcting the confidence of CPAR classification rules[22] probabilities in Naïve Bayes models[23]. White-box approaches have not been further developed in recent years, being superseded by in-processing methods hence, black-box methods are preferred for post-processing.

Black-box approaches modify the models' predictions. Examples of the black-box approach aim at keeping proportionality of decisions among protected versus unprotected groups by promoting or demoting predictions close to the decision boundary[24], by differentiating the decision boundary itself over groups[25], or by wrapping fair classifier on top of a black-box base classifier[26], Equalized odds postprocessing[27] solves a linear program to find probabilities with which to change output labels. Calibrated equalized odds postprocessing[28] optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective. Reject option classification[29] gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

3 Conclusion

Artificial Intelligence is more efficient than humans in solving tasks, and as time progress so does its ability to solve more complicated problems. In order to apply AI in real-world scenarios, Artificial Intelligence bias needs to be removed so that no group is affected negatively. This paper has classified bias into 3 major types and has described how bias may creep inside the model.It has spoken about three stages where bias can

Table 3 Main differences between black box and white box post processing methods.

	White box methods	Black box methods
What do they consist of?	They consist of altering the model internals and they post-process the classification model once it has learned from data.	They modify the predictions of the model.
Preference in usage.	They are not developed and in-processing methods are preferred.	They are preferred for post-processing as research has been done.
What are some of its features?	It consists of detecting problems in the code and product.	It consists of detecting problems in the features and performance of the model.

be mitigated. The goal of the paper was to spread awareness on the latest findings on the topic and to familiarize people with the latest findings For future research, readers can read the papers referenced as well as see which methods have been developed in the paper such as IBMs AI 360. The field of AI bias is constantly being researched and mitigation techniques are being developed further.

References

1 V. Shashkina, *What is AI bias really, and how can you combat it*, https://itrexgroup.com/blog/ai-bias-definition-types-examples-debiasing-strategies/#header, Internet]? Available from:.

2 W. Dieterich and M. B. Ph.D. Christina Mendoza, *Ph.D*, Demonstrating Accuracy Equity and Predictive Parity, COMPAS Risk Scales, p. 1–3.

3 B. Friedman and H. Nissenbaum, *ACM Transactions on Information Systems*, **14**, 330–347.

4 C. Dilmegani, *Bias in AI: What it is, Types, Examples 6 Ways to Fix it in 2022*, https://r21esearch.aimultiple.com/ai-bias/, Internet];[about 2 screens]. Available from:.

5 D. Roselli, J. Matthews and N. Talagala, Companion Proceedings of The 2019 World Wide Web Conference.

6 M. Del Balso and J. Hermann, *Uber Engineering*.

7 B. D., Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Nova Scotia, Canada.

8 M. Ribeiro, S. Singh and C. Guestrin, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY.

9 J. Buolamwini and T. Gebru, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY.

10 E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl and V. M, *WIREs Data Mining and Knowledge Discovery*, **10**, year.

11 B. Luong, S. Ruggieri and F. Turini, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, p. – 11.

12 F. Kamiran and T. Calders, 2009 2nd International Conference on Computer, Control and Communication.

13 T. Calders, F. Kamiran and M. Pechenizkiy, 2009 IEEE International Conference on Data Mining Workshops.

14 V. Singh and C. Hofenbitzer, Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

15 F. Calmon, D. Wei, B. Vinzamuri, K. Ramamurthy and K. Varshney, *IEEE Journal of Selected Topics in Signal Processing*, **12**, 1106–1119.

16 R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork, *Proceedings of the 30th International Conference on International Conference on Machine Learning*, **28**, –325– –333.

17 M. Feldman, S. Friedler, J. Moeller, C. Scheidegger and S. Venkatasubramanian, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

18 G. Smith and I. Rustagi, *Mitigating Bias in Artificial Intelligence*.

19 R. Schwartz, L. Down, A. Jonas and E. Tabassi, *A Proposal for Identifying and Managing Bias in Artificial Intelligence*.

20 P. Lohia, K. Natesan Ramamurthy, M. Bhide, D. Saha, K. Varshney and R. Puri, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP.

21 J. Silberg and J. Manyika, *Tackling bias in artificial intelligence (and in humans*.

22 S. Ruggieri, D. Pedreschi and F. Turini, *ACM Transactions on Knowledge Discovery from Data*, **4**, 1–40.

23 T. Calders and S. Verwer, *Data Mining and Knowledge Discovery*, **21**, 277–292.

24 F. Kamiran, S. Mansha, A. Karim and X. Zhang, *Information Sciences*, **425**, 18–33.

25 P. Hardt and S. Supervised Learning.

26 B. Agarwal, L. Dudík and Wallach, *A Reductions Approach to Fair Classification*.

27 M. Hardt, E. Price, E. Price and N. Srebro, *Equality of Opportunity in Supervised Learning*.

28 G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg and K. Q. Weinberger, *On Fairness and Calibration*.

29 N. T. L. a, *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms [Internet*, https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best\-practices-and-policies-to-reduce-consumer-harms/, 2022 [cited 7 June 2022]. Available from:.