Ruqayyah Muse

Dr. Murray Patterson

HON Data Science 4810

01 March 2023

Bias in Artificial Intelligence

In a simple machine learning life cycle, a business problem is defined, some data is gathered, a model is trained and tested on this data, and then if it is deemed accurate at making predictions, this model is then deployed in the real world to solve the target task. Even after deployment, this model can continue to grow in accuracy by learning from new data thus making it an intelligent agent. Models and agents built using this process have become an integral part of our lives and have helped to automate processes for companies and consumers. Siri, Google Assistant, self-driving cars, and automated hiring systems are all examples of intelligent agents that can affect our daily lives. These agents can be seen as impartial, unbiased systems that will help to increase equality and fairness when it comes to areas such as hiring or loan approval. However, human biases can still find their way into these agents, resulting in potentially harmful systems that not only perpetuate human injustices but at times exacerbate them.

By definition, bias in machine learning and artificial intelligence (AI) agents is a model's tendency to predict or make a decision that is for or against one group in a way that is considered unfair (Ntoutsi et al. 3). These biases reflect human and societal norms and are usually introduced during the development process of a model, such as in the data collection, model selection, or model evaluation steps. Even in a situation where the process is perfectly free of bias, the fact that we live in an imperfect world with a history of discrimination and inequality

means that the data we use to train models will probably always include some form of historical bias.

Historical bias is defined as bias arising "if the world *as it is* or *was* leads to a model that produces harmful outcomes" (Guttag and Suresh 4). From the history of just the United States alone, we can argue that any amount of data gathered over time will reflect the injustice and bias of the people who lived in that period. Even in the modern day, there are still a lot of internalized and rampant forms of biases, such as gender and racial biases, that influence the kind of data available for models to learn from. In this paper, I will: explain further what gender and racial biases are and share real-world cases where AI models exhibited these biases; analyze sources of bias and the impact of historical bias in the model development process; and critique some proposed methods for mitigating the effects of bias in deployed models.

GENDER AND RACIAL BIAS:

Gender bias is defined as the "preference or prejudice toward one gender over the other" (Sun et al. 1). It is found primarily in Natural Language Processing (NLP) models which can return gender-biased predictions that reinforce stereotypes or unfairly give preference to one gender over the other in applications such as resume filtering or language translation. Some notable cases of gender bias in real-world applications occurred with Amazon's resume screening program that passed over female candidates in preference of male ones and was found to be discriminatory in 2015 (Caliskan). Another case where gender bias occurred was with Google Translate's algorithm which systematically translated statements from gender-neutral languages into stereotypical gendered statements in English (Caliskan). GPT-3, a predecessor of the infamous ChatGPT, is another model that was also found to be gender biased as it generally classified occupations 83% of the time as being male, or suggested appearance-based adjectives

to describe women while adjectives for men were from a larger variety of words (Brown et al. 36-37). After the discovery of gender bias in these applications, the companies behind these models took action to reduce the levels of bias in the outputs, or in some cases, stopped using the model altogether.

Similarly, racial bias is the preference or prejudice of one race over another and has been a part of human history for hundreds of centuries. As a result, this bias has crept into some machine learning models and algorithms such as those implemented within the criminal justice or financial systems. One case where racial bias might be inherent is with risk assessment algorithms used within the criminal justice system. These risk scores can be used to determine bail amounts and are at times provided to judges to consider during sentencing (Angwin et al.). However, the algorithms behind these assessments have been found to issue higher risk scores for people of color compared to white offenders raising concerns about the legality of such systems (Angwin et al). Other cases where racial bias might inadvertently be found are with facial recognition systems. Many of these algorithms were found to be less accurate in classifying the faces of darker-skinned people, particularly darker-skinned women (Guttag and Suresh 6). However, a lot of times racial and gender bias is not intentional and is simply a product of models learning from data that includes biased classifications, or from data that is not representative enough of minority groups.

THE SOURCES OF BIAS:

Frequently, the source of bias in models is attributed to the data that models are trained on. However, data itself is not always necessarily biased; the process of collecting it and the way it is applied by the developers of a model determines how much effect the bias in the data, if any, will have on the final outputs of the model (Guttag and Suresh 1). It is important to realize that the

entire development process from collecting data, choosing a model to learn from this data, evaluating the accuracy of the chosen model, to deploying the model are all avenues where bias can be introduced.

The first point for bias to be introduced is in the collection process of data. The way data is gathered and the level of representation within the data have a significant impact on the assumptions that a model ends up learning. For instance, text data that are drawn primarily from the 1950s might reflect historical racial and gender biases which would result in a model that produces racist and sexist content. In our current world, such an output would be unacceptable, but the information available to the model to learn from was not gathered in a way to reflect modern views. Similarly, a facial recognition algorithm that is trained primarily on a dataset of lighter-skinned faces will perform poorly on darker-skinned faces since they are not properly represented within the data. This source of bias can be easily remedied by developers being more careful and deliberate in gathering representative and less biased data.

Another source of biased predictions can be a result of the type of model that is implemented. One example is that of a general NLP model trained on all Twitter data inaccurately interpreting emojis or hashtags that have gang-related meanings (Guttag and Suresh 5). In a general setting, such an NLP model might be accurate in interpreting Tweets; however, if there are underlying groups or patterns within the data, another model might be more fitting to account for these subsets that don't follow the general form of the data (Guttag and Suresh 5). This is the case with many datasets: there are always subsets or underlying patterns within the data that are not consistent with the general mapping algorithm that one model might implement. To avoid skewed predictions in this case, developers would need to choose a more detailed

model that can understand these subsets and not be biased towards the more general, larger category in the dataset.

A third source of bias is evaluation bias which can occur when different models are compared using one performance measure that doesn't properly represent the accuracy of each model. For instance, in the case of facial recognition models, taking the average accuracy of different models might suggest that one model is more accurate than others in recognizing faces. However, that performance measure does not consider that the algorithm might be underperforming with subgroups, such as darker-skinned women (Guttag and Suresh 6). There might be another model that is overall slightly less accurate but consistently better at recognizing different subgroups, making it a more suitable model to deploy. Because of this, evaluation bias can be more difficult to remedy, as it is easier to have one single arbitrary value to judge between models. Nonetheless, for some problems, it is important to evaluate models using more detailed performance measures that better reflect what a truly accurate prediction should be.

MITIGATING BIAS

Researchers within the field of AI and machine learning have suggested various methods for mitigating bias such as creating synthetic data, using data augmentation to make more representative data, bias fine-tuning models before giving them biased data, or even spot-training noted biases after the deployment of a model. In this section, I explore the advantages of these methods as well as their limitations.

Using Synthetic Data. Synthetic data is defined as computer-generated information that can be used in place of personally identifiable data, such as race, religion, and sexuality (Martineau). It can be an easier, simpler solution to mitigating bias since it avoids legal restrictions that are

aimed to protect the privacy of individuals, and it can also be used to generate diversified data. Despite this, synthetic data might not always be the best method for mitigating bias. To produce synthetic data, models must still be trained on real and original data, so the synthetic data could still end up reflecting the bias of the real world (Corrêa). On the other hand, if they are not trained on real data, then the created data would not be representative of the real world, and thus models trained on that synthetic data wouldn't be able to make predictions that fit the world as it is.

This paradox makes it difficult to adopt synthetic data in all cases. But it could still be used on a smaller scale to fill in real datasets that are not very representative. For instance, to solve the issue of a facial recognition dataset that is composed primarily of lighter-skinned faces, we could train a model on the subset of available darker-skinned face instances and produce synthetic instances that could be added to the original dataset. By augmenting the original dataset with synthetic data, the representation imbalance within the original dataset would be alleviated and would hopefully lead to models that are better at recognizing any face, regardless of skin tone.

Using Data Augmentation in NLP models. Another method that could be applied specifically with NLP models is data augmentation to reduce gender bias. In a dataset that is biased towards one gender, we can create a duplicate dataset with the same data after swapping the gender of the speaker. For instance, if the statement 'He is a doctor' occurred in the original dataset, then we'd add a sentence with the subject's gender being female to the new dataset: 'She is a doctor'. Likewise, a statement such as 'She is a baker' would have an equivalent 'He is a baker'. After swapping the genders for all entries, the two datasets would then be combined, resulting in a balanced dataset that would make the subject of a sentence have a 50% chance of being female

or male (Sun et al. 5). A model trained on this balanced dataset would then be equally likely to predict that the subject of a sentence such as '___ is a nurse' to be male or female, thus removing gender bias.

The limitation of this method is usually the cost. Performing data augmentation can be time-consuming and expensive when it comes to creating and storing additional data (Sun et al. 5). A developer would have to make a copy of every given dataset with gender-swapped subjects and then train the model on the union of these two datasets. This could be almost impossible to do if the original dataset is made up of terabytes or petabytes of data. Another limitation is that some sentences can be nonsensical, such as gender-swapping 'She gave birth' to 'He gave birth' (Sun et al. 5). Without accounting for exceptions such as this, the model might end up making predictions for the subject that do not make sense in the real world. As a result, while data augmentation can be a useful solution, the cost of implementing it makes it infeasible for many models.

Removing Personal Identifying Information. Another solution that is at times used to mitigate racial or other forms of identity discrimination is removing personal data such as race, religion, and sexuality from datasets. The idea is that removing such data stops a model from learning correlations that are based on attributes that are considered discriminatory, such as a model learning that there is a correlation between the race of an applicant and their success at receiving a loan. However, even by removing these protected identifying features, models can still find hidden correlations that we might not ordinarily notice (Corrêa). For instance, a model that does not have information about race or ethnicity could learn that people from a certain zip code are typically denied loans. That zip code could then prove to have a majority that is of one ethnic

group. Despite the ethnicity of the people not being included in the dataset, something as innocent-seeming as an address could become a proxy for that information.

A real-world example of this is the Northpointe Risk Assessment score that is used by some jurisdictions in the criminal justice system. The risk score is determined by a person's response to a survey of questions such as "Was one of your parents ever sent to jail or prison?", "How many of your friends/acquaintances are taking drugs illegally?" or "How often did you get in fights while at school?" (Angwin et al.). Importantly, this risk assessment survey does not ask about a person's race. However, for a Black person who comes from a disadvantaged background, their answers to these questions will likely be in the affirmative because of the environment that they live in. Even if this is their first confrontation with the law and the model does not explicitly know their race, the percentage of black people receiving a higher risk score has been found to be greater compared to white people (Angwin et al). This could potentially be a result of historical correlations between disadvantaged backgrounds and race and might be a more difficult bias to combat since it is in a way reflecting the reality of the world.

Bias Fine-Tuning and Targeted Retraining. Two other methods for combating bias include bias fine-tuning and targeted model re-training. Bias fine-tuning is when a model is first trained on a related, less biased dataset to ensure that the model has minimal bias before it is trained on the biased dataset for the target task (Sun et al. 6). On the other hand, target model re-training occurs after a model has been deployed. Basically, when bias is noticed in the model's output, the model is retrained on new, smaller datasets designed to specifically target the noted bias (Saunders and Ullmann). Both of these methods have not been extensively applied or researched but could prove to be promising solutions to explore for mitigating bias.

Conclusion

Based on this discussion exploring different types of biases, investigating the avenues where bias is introduced, and analyzing some proposed methods for mitigating bias, it's easy to see how bias can unexpectedly play a role in exacerbating real-world injustices, and how there is also no simple method for mitigating it within AI models. A lot of issues that come with trying to remove bias in AI stem from the fact that the world has always contained prejudices and inequalities that end up affecting the entire process of building AI models. Hence, by virtue of living in an imperfect world, it is possible that models will also always be imperfect and contain some degree of unfairness. However, steps can still be taken to reduce that bias which can hopefully lead to a more equitable and fairer world in the future.

Works Cited

Angwin, Julia, et al. “Machine Bias.” *ProPublica*, 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Brown, Tom B., et al. “Language Models Are Few-Shot Learners.” *ArXiv.org*, ArXiv, 22 July 2020, https://arxiv.org/abs/2005.14165.

Caliskan, Aylin. “Detecting and Mitigating Bias in Natural Language Processing.” *Brookings*, The Brookings Institution, 9 Mar. 2022, https://www.brookings.edu/research/detecting-and-mitigating-bias-in-natural-language-processing/.

Corrêa, Ana Maria. “Is the Fate of Equality Synthetic?” *Centre for IT & IP Law*, Ku Leuven CiTiP, 21 Apr. 2022, https://www.law.kuleuven.be/citip/blog/is-the-fate-of-equality-synthetic/.

Martineau, Kim. “What Is Synthetic Data?” *IBM Research Blog*, IBM, 8 Feb. 2023, https://research.ibm.com/blog/what-is-synthetic-data.

Ntoutsi, Eirini, et al. “Bias in Data-Driven Artificial Intelligence Systems—an Introductory Survey.” *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2020, p. 3., https://doi.org/10.1002/widm.1356.

Sun, Tony, et al. “Mitigating Gender Bias in Natural Language Processing: Literature Review.” *ArXiv.org*, ArXiv, 21 June 2019, https://arxiv.org/abs/1906.08976.

Suresh, Harini, and John V Guttag. “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle.” *ArXiv.org*, ArXiv, 1 Dec. 2021, https://arxiv.org/abs/1901.10002.

Ullmann, Stefanie, and Danielle Saunders. “Online Translators Are Sexist – Here's How We Gave Them a Little Gender Sensitivity Training.” *The Conversation*, 15 Sept. 2022, https://theconversation.com/online-translators-are-sexist-heres-how-we-gave-them-a-little-gender-sensitivity-training-157846.