Addressing Algorithmic Bias in Artificial Intelligence

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Riley Hartung

Spring, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Bryn E. Seabrook, Department of Engineering and Society

STS Research Paper

Introduction

Most facial recognition algorithms perform more accurately on White faces than they do Asian, African American, or Native American faces (NIST, 2019). This was the finding from a study conducted by the National Institute of Standards and Technology (NIST), an organization that is part of the U.S Department of Commerce and holds inclusivity as one of their core values. Their study involved almost 200 algorithms from 99 developers, which made up most of the industry at the time. The testing was done on 18 million images sourced from the State Department, Homeland Security, and the FBI. After comparing the accuracy of the algorithms' facial matching capabilities, NIST concluded that the algorithms are anywhere from 10 to 100 times worse at matching African American than White faces.

This is just one example of how discrimination is manifested in artificial intelligence (AI). An AI research team from IBM defines discrimination in machine learning as, "when a given individual different from another only in 'protected attributes' (e.g., age, gender, race, etc.) receives a different decision outcome from a given machine learning (ML) model as compared to the other individual." (Aggarwal et al., 2019). Because AI is present in many aspects of everyday life (Marr, 2019), these biases are found almost everywhere. This paper examines the question of how algorithmic bias in artificial intelligence can be reduced.

Methods

As use of AI grows in the world, so too does the amount of algorithmic discrimination. Many companies are adopting groundbreaking ML techniques without considering how they might impact their user base. These recent developments raise the question, how can algorithmic

bias in AI be reduced to prevent further harms to minority groups? Keywords related to this paper include algorithmic bias, artificial intelligence, machine learning, and discrimination.

The research and findings of this paper are organized into two sections. First is a discourse analysis of anecdotes from those who have been affected by discriminatory bias in AI. This details how bias is propagated by AI in different fields and the effect it has on the people it harms. However, it is not a comprehensive overview of all the potential sites that machine learning, but rather an in-depth look at a few key examples of it occurring.

The next section is a documentary analysis, and the sources are primarily research papers written by scholars of AI. This half of the methods catalogs the different methods of detecting, preventing, and removing discrimination from ML models that were found within the sources and is arranged chronologically according to the lifecycle of a model. There is discussion of how to properly vet data for training, how to conduct that training, methods for identifying discrimination with the training results, and resources for monitoring the behavior of a model; all in ways that reduce the amount of bias produced by that model.

What is Machine Learning?

Machine learning is a subset of AI. While ML is a powerful tool for solving problems, it is not necessary for every application. An artificially intelligent solution is best suited when the rules for obtaining the answer cannot be determined or coded by a human, or when the scale requires the process to be automated (AWS, 2016). For example, determining whether an email is spam would be an appropriate task for ML, as there are many variables that contribute and interact and coding in all of them is difficult. A person may be able to manually decide which emails are spam, but it would become very tedious if they had to do this for hundreds or even

thousands of emails (AWS, 2016). These types of complex or high-scale problems appear in every sector, so it is not surprising that the healthcare, retail, manufacturing, finance, and technology industries are all utilizing AI (CSU Global, 2021).

ML models can be split into two categories: white-box and black-box. The former, also known as explainable artificial intelligence (XAI), is made up of models that have easily interpreted results (by experts in the field), and the latter is composed of those that are hard to be explained and understood from a mathematical perspective (Loyola-González, 2019). Typically, in situations where one type performs well, the other type performs poorly. While this is an important consideration when constructing a model, sometimes the situation requires a white-box solution. The Equal Credit Opportunity Act requires that denial of credit must not have vague or indefinite reasoning, so black-box models cannot be used for this purpose. Additionally, systematic bias, such as discrimination based on protected features (age, gender, race, etc.), is very hard to detect for end users. Transparency in how algorithms operate is necessary for identifying discrimination and maintaining accountability (Alufaisan et al., 2016).

There also exists two different basic types of training for these models: supervised and unsupervised learning. In supervised learning, each training data point is labeled with the target, or the desired prediction for that point, and it must have labeled features, which are the attributes that help to identify that target (AWS, 2016). This data is fed to the model in a process called training. After training, more data is necessary to evaluate the accuracy of the model. The ML model receives data, determines the target based on the features of that data, and then its performance is rated by the ratio of correctly chosen targets. Unsupervised learning, by comparison, does not require a human to classify or label data (IBM, n.d.). In this type of learning, the model clusters the data based upon or features it identifies itself. This can be useful

for finding hidden patterns between data points, but due to its hands-off nature, the results can sometimes be undesired or hard to explain.

The Social Construction of Artificial Intelligence

AI is ubiquitous in human-computer interactions. Social media, navigation apps, smart home devices are just a few examples of where AI can be found in everyday life (Marr, 2019). This technology is also integrated into more personally impactful systems. Many screening processes involve use of ML, such as college admissions and vetting candidates for job interviews. With such an established place in modern society, ML algorithms hold a lot of influence in the day-to-day and the larger experiences of people. When discriminatory biases are present in these algorithms, they disadvantage groups of people from attending schools, gaining employment, receiving credit, and receiving accurate medical diagnoses. These discrepancies are compounded for marginalized communities, who already face discrimination from countless other systems.

The social construction of technology (SCOT) will be used to analyze this research question. This framework was first formalized in "The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other" from 1987 by Trevor Pinch and Wiebe Bijker (Klein & Kleinman, 2002). SCOT specifically analyzes the process of creating technological solutions to problems, so it will be useful in examining the building stages of ML algorithms.

It is important to note that this framework makes some assumptions about how the construction of technologies is performed. One of these assumptions is that multiple different social groups are present for the planning and building of a given technology. This is not always

true, and in some cases of ML algorithms, the construction is done by a single developer or data scientist. SCOT also assumes that these groups are equal in their representation and power (Klein & Kleinman, 2002). This claim in particular is rarely true in practice as groups in industry are regularly overruled on the basis of numbers, seniority, experience, position, etc.

Assumptions about the technological creation process manifest as systems that are discriminatory and do not equally serve the public.

Results and Discussion

1. Introduction

If you fall into a minority group and have ever applied to a job site (Gilbert, 2023), applied for credit (Andrews, 2021), messaged a chat bot (Schwartz, 2019), used facial recognition (NIST, 2019), or received treatment at a hospital (Ledford, 2019), you've probably experienced algorithmic bias. Alarming amounts of discriminatory bias has been found in all of these applications and more. Unfortunately, there is no single solution for eliminating discrimination from AI. This paper addresses non-technical solutions in the first half of the results section which includes raising awareness of the ways in which algorithms are perpetuating bias, recognizing accounts of bias, and giving disadvantaged groups platforms to represent themselves and take part in the design of such technologies. Technical solutions are in the latter half of the results and are ordered by the development stage that they take place in (vetting data, conducting training, and monitoring output).

2. Non-technical Solutions to Bias

2.1 Dr. Buolamwini's "Coded Gaze"

Dr. Buolamwini holds a PhD in Media Arts and Sciences from the MIT and now conducts research there. She is seen as an expert in the field of AI and has served on several

panels to discuss how it is perpetuating discrimination. She is also well known for founding the Algorithmic Justice League (AJL), which aims to increase public awareness about the impacts and implications of AI. One of the reasons that Dr. Buolamwini has dedicated so much of her life to the study of AI is because of the impact it has had on her herself.

Several years ago, the then master's student was working on a project she called the Aspire Mirror (Buolamwini, 2018). The purpose of the mirror was to identify your face and overlay another face over it. If you wanted to feel powerful, you could choose to cast a lion over your face and see it move in real time. Before she could change how she looked, she first had to program the device to recognize her face when it appeared in the camera. To her dismay, the program was totally unable to identify her. It was not until she put on a white mask that the device was able to register her presence. Aside from being humiliating in itself, this act also echoed the expectation for minorities to put on a cultural mask every day. Dr. Buolamwini recollects the experience not as an isolated event but as part of an ongoing trend. Biases in technology like this only pile upon the discrimination that minorities face in nearly every aspect of life.

The failure of the Aspire Mirror was the discovery that started Dr. Buolamwini down her road to algorithmic justice. She saw that technologies were being created and implemented without consideration of their adverse effects and coined the term "coded gaze" to mean algorithmic biases that perpetuate discrimination and social inequities. Dr. Buolamwini is just one of many individuals who have experienced exclusion at the hands of AI. She works in hopes that more diverse perspectives will be included in the conversations surrounding AI and ML and reduce the exclusionary properties of many modern "smart" technologies.

While the repercussions of a facial recognition error in Dr. Buolamwini's project may not be dire, there are many other applications of facial recognition for which there would be severe consequences if a similar problem were to occur. For example, facial recognition technology has been used by law enforcement in the U.S. to identify suspects (Taylor et al., 2020). It was found that this tech misidentified people of color, especially women of color, at much higher rates than other groups. These unintentional results are often due to the data that the model is trained on. In the 2019 study referenced in the introduction of this paper, researchers found that of the 189 algorithms tested, many gave false positive matches for Black faces (NIST). While the research did not examine the cause of the inaccuracies, one hypothesis was it is due to, "the data used to train it." (NIST, 2019).

It is important to recognize that these biases occur in part because of lack of representation in the teams building and testing the algorithms. One of the assumptions of the SCOT framework is that there are different social groups present in the construction of technology (Klein & Kleinman, 2002). The tech industry is overwhelmingly White and male, and according to a 2021 study, less than 5% of U.S. computer scientists were Black or African American (*Computer Scientist Demographics and Statistics [2023]*, 2021). This leads to development teams that are not representative of the population and are more prone to excluding minorities, racial and other. Preventing discriminatory bias in AI will require diverse teams at every step of the process.

2.2 Employment Discrimination

AI is also widely used in the job application process. Applicant screening algorithms are built by companies like Workday to filter out those who do not meet the position's requirements. However, these algorithms are built by people and trained by data that may have biases

themselves. In February 2023, Derek Mobley, a Black man from California, filed a complaint against Workday concerning these biases (Gilbert). He sought to represent applicants belonging to protected groups that had been discriminated against by the screening process and denied equal employment opportunity. Despite having a bachelor's degree in finance and an associate's degree in network systems administration, Mobley says he has been denied from over 80 positions since 2018, all of which used Workday's screening algorithm. It his concern that his status as a Black man over the age of 40 has led to his applications being flagged as less qualified than other applicants who are in a racial majority and are younger. Workday responded that Mobley's claim of algorithmic bias is unfounded, as they, '"engage in a risk-based review process throughout our product lifecycle to help mitigate any unintended consequences, as well as extensive legal reviews to help ensure compliance with regulations."' (Gilbert, 2023). For any reduction in algorithmic discrimination to happen, corporations need to take responsibility for their algorithms and conduct robust testing. This attitude of denying all claims of bias is not conducive towards resolving the problem at hand.

2.3 Using Diversity to Prevent Exclusion

The principal way in which exclusion is allowed to occur is through lack of diversity and representation. For AI algorithms, a diverse group is needed both within the training set and in the team responsible for building the technology. SCOT assumes that all "relevant social groups" are present and have equal say in the design process (Klein & Kleinman, 2002). However, for algorithms like facial recognition used in identifying suspects, the relevant social groups are those of the entire country. The data that models are trained on need to be representative of the population it will be used for.

Due to its rising popularity, companies across all sectors are increasing their stakes in AI. These additions are often made without considering the resources and infrastructure necessary for preventing bias. As discussed above, a significant amount of investment is required to develop fair and inclusive models. To work towards a world with more equitable AI, actors need to recognize that ML can have widespread, harmful effects that can also exacerbate inequalities experienced by minorities. The baseline requirements to prevent releasing discriminatory algorithms include a diverse training set, a diverse development/oversight team, and comprehensive bias testing methods and resources. Without these, companies should not be implementing potentially harmful AI systems.

3. Technical Solutions to Bias

3.1 Vetting Data

In an attempt to create a tool for eliminating bias from training data, three researchers at North Carolina State University have developed Fair-SMOTE, an algorithm that removes biased labels and rebalances the data distribution (Chakraborty et al., 2021). Fair-SMOTE is differentiated from previously published methods because 1. when tested against another algorithm, this new method was 220% faster, and 2. this method was also more accurate in terms of recall and F1. In ML, recall is the percentage of samples that a model correctly identifies, and F1 is a statistic that combines the recall and precision of a model.

Fair-SMOTE rebalances data by generating new data points for subgroups until they all are of equal size. First, a random point from a subgroup is selected to be the parent point. An algorithm is used to find the two nearest neighbors to the point, and the attributes for the new point are calculated using the attributes of the parent and neighbor points. This process is repeated until all subgroups are of equal size. Associations between data points are preserved

through extrapolating from the neighboring points. Extrapolating helps to ensure the model develops the same pattern recognition during training.

Fair-SMOTE effectively reduces model bias and increases fairness. For this study, the researchers used four metrics for measuring bias. Two of them relate to the difference of false and true positive rates between privileged and underprivileged groups, while the other two measure the difference in probability of a privileged group or unprivileged group receiving a favorable prediction. "Fair-SMOTE performs similar or better than two state of the art bias mitigation algorithms in case of fairness and consistently gives higher recall and F1 score." (Chakraborty et al., 2021). In short, Fair-SMOTE reduces model bias without losing accuracy or precision.

3.2 Conducting Training

Artificially intelligent systems produce output using the associations they have built during training. If an AI has developed associations that cause undesirable biases, it must be retrained to alter or remove those biases. Researchers from the Korea Advanced Institute of Science & Technology in South Korea have designed an algorithm to avoid unwanted associations which they call Learning Not to Learn (LNTL) (Kim et al., 2019). Their method operates using regularization, which is a form of regression that relates to how much a model utilizes noise, or extraneous data, to make a classification (Gupta, 2017). Reducing the regularization can prevent a model from overfitting, which is making associations using noise that do not generalize to other data. The researchers report that "the proposed regularization term minimizes the detrimental effects of bias in the data" and "the network was able to learn more informative features for classification" (Kim et al., 2019). They found that their specially trained model performed better than compared models in most experiments.

Despite the success of the Korean research team, successive papers have observed worse results after applying debiasing. A team from the University of Campinas, Brazil applied LNTL to a model designed to classify skin lesions and found that performance suffered (Bissoto et al., 2020). Debiasing is also less effective in certain applications of AI. Natural language processing (NLP) is a type of ML that allows computers to understand word meanings (IBM, n.d.). It is used in smart assistants, online translators, and chatbots. Instances of discriminatory bias are hard to avoid in NLP, as illustrated by Microsoft's Twitter bot, which tweeted a large number of antisemitic, racist, and other generally offensive comments during its short demo (Schwartz, 2019). The problem with word embedding debiasing is it hurts the overall effectives of the NLP model (Caliskan, 2021). For example, removing the gender association with certain words would change how effectively the model reports occupational gender statistics. These results suggest that more study is necessary in debiasing for complex analyses like images and NLP.

The best resource for studying the relationship between word embeddings and NLP outputs is the training data for the model and the values of the word embeddings themselves, however, these are not available to the public due to data privacy laws (Caliskan, 2021). Legislative reform is necessary for effective debiasing research to take place. There also needs to be standards regarding AI testing and what is fit to go to market, regarding the presence of bias and discrimination.

3.3 Monitoring Output

Researchers at the University of Texas published a paper in 2016 in which they developed two methods to measure the discrimination of black-box models (Alufaisan et al., 2016). As previously mentioned, the results of black-box models are not explained to end users, so it is vital to have metrics for measuring bias in them. The researchers used 80% disparate

impact (DI) as the minimum amount to constitute discrimination, which is the rule recommended by the U.S Equal Employment Opportunity Commission (EEOC). "In U.S law, disparate impact is the indirect or intentional discrimination for different groups based on their protected features." (Alufaisan et al., 2016). DI was calculated by dividing the probability of a good result (e.g., credit approval) given that the subject was in the minority by the probability of a good result given that the person was in the majority.

The first of these techniques is named *Machine Learning extended lift* (ML*lift*). ML*lift* builds upon the predecessor *elift* method, which measures discrimination between two independent itemsets. However, ML*lift* is capable of measuring discrimination in data containing any number of features (Alufaisan et al., 2016). This technique is performed using ML models that can compute the probability that data points belong to specified classes. Several different models were used, and they all gave similar results, but testing with a support vector machine model (SVM*lift*) gave the most consistent results so it was used as the default ML*lift* model.

The second technique developed was *Feature-based Targeted Sampling* (FTS) (Alufaisan et al., 2016). FTS builds artificial data collected from the value of features from two different groups. The first group is made from data that was assigned a good result by the model being tested, and the second group are instances of a subgroup with a protected attribute. Due to the selection of group members, only a small number of the sample should be classified as bad. If there is a disproportionately high number of bad classifications within the selected data, then the model being tested is biased against the protected subgroup. Discrimination within the subgroup is calculated by taking the sum of outputs of a Random Forrest model (RF) that takes the sample data as input. A discrimination score can be computed by taking the difference of the score given

by the protected subgroup and the score of another sample comprised of members from the majority group.

The research team concluded that both ML*ift* and FTS were effective discrimination measures for black-box classifiers. When the classifier was tested on unbiased data, unbiased results were received. Upon providing discriminative datasets, discriminative results were detected by the two measures and those amount was verified by calculating DI.

4. Limitations and Future Research

While the methods described above are useful in reducing algorithmic bias in some applications, they are not generalizable to every use of AI. Incompatibilities can be caused by the format of the input or output expected, the type of algorithm being used, or in the case of NLP, modifying results in worse performance. For example, Fair-SMOTE relies on the ability of creating new data points, but for models that rely on images for their inputs, Fair-SMOTE cannot generate new data points with its existing point generation code.

Future research should include bias reduction methods for AI techniques that were not covered in this paper. There should also be an investigation on what steps are required for enacting legislation concerning AI standards. As NLP is a quickly growing field of AI, it is vital that word embeddings and training data used for these models are released to researchers. The U.S. federal government should also set in place standards for testing bias, and how much bias is allowed for commercial use of AI.

Conclusion

Effectively reducing discriminatory bias in AI will require both technical and nontechnical solutions. The first step to fixing the problem is recognizing where it exists. All

relevant social groups, including minorities, need to be included in the training and development process for ML models so that they are not built solely for the majority. Additionally, technical solutions such as LNTL and Fair-SMOTE should be used in modifying data, debiasing, and output monitoring. Tools like these should be seen as supplementary to having a diverse and representative development team, not as a replacement. Ultimately, companies employing AI in their products need to invest more resources in designing and testing their algorithms to avoid further marginalizing minority groups.

References

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., & Saha, D. (2019). Black box fairness testing of machine learning models. *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 625–635. https://doi.org/10.1145/3338906.3338937

Alufaisan, Y., Kantarcioglu, M., & Zhou, Y. (2016). Detecting Discrimination in a Black-Box Classifier. *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, 329–338. https://doi.org/10.1109/CIC.2016.051

Amazon Machine Learning—Developer Guide. (2016).

Andrews, E. L. (2021, August 6). *How Flawed Data Aggravates Inequality in Credit*. Stanford HAI. https://hai.stanford.edu/news/how-flawed-data-aggravates-inequality-credit

Bissoto, A., Valle, E., & Avila, S. (2020). Debiasing Skin Lesion Datasets and Models? Not So Fast. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3192–3201. https://doi.org/10.1109/CVPRW50498.2020.00378

Buolamwini, J. (2018, June 26). *Fighting the "coded gaze."* Ford Foundation. https://www.fordfoundation.org/news-and-stories/stories/posts/fighting-the-coded-gaze/

Buolamwini, J. (2022). *Facing the Coded Gaze with Evocative Audits and Algorithmic Audits* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/143396

Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in Machine Learning Software: Why? How? What to do? *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 429–440. https://doi.org/10.1145/3468264.3468537

Christie, T., Benjamin, R., & Raji, D. (2020, June 19). Black Lives Matter Protests Shine Light On Facial Recognition Problems. *Science Friday*. https://www.sciencefriday.com/segments/ai-equity/

Computer Scientist Demographics and Statistics. (2021, January 29). https://www.zippia.com/computer-scientist-jobs/demographics/

Gilbert, A. (2023, February 22). *Workday AI Biased Against Black, Older Applicants, Suit Says (1)*. Bloomberg Law. https://news.bloomberglaw.com/daily-labor-report/workday-ai-biased-against-black-disabled-applicants-suit-says

Gupta, P. (2017, November 16). *Regularization in Machine Learning*. Medium. https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

Hamilton, I. A. (2018, October 13). *Why it's totally unsurprising that Amazon's recruitment AI was biased against women*. Business Insider. https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10

IBM. (n.d.). *What is Natural Language Processing? | IBM*. Retrieved March 22, 2023, from https://www.ibm.com/topics/natural-language-processing

Kim, B., Kim, H., Kim, K., Kim, S., & Kim, J. (2019). *Learning Not to Learn: Training Deep Neural Networks with Biased Data* (arXiv:1812.10352). arXiv. http://arxiv.org/abs/1812.10352

Klein, H. K., & Kleinman, D. L. (2002). The Social Construction of Technology: Structural Considerations. *Science, Technology, & Human Values*, *27*(1), 28–52. https://doi.org/10.1177/016224390202700102

Lasorsa, C. (n.d.). *An Introduction to Automated Data Labeling*. Retrieved February 5, 2023, from https://www.superb-ai.com/blog/an-introduction-to-automated-data-labeling

Loyola-González, O. (2019). Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access*, *7*, 154096–154113. https://doi.org/10.1109/ACCESS.2019.2949286

Marr, B. (2019, December 16). *The 10 Best Examples Of How AI Is Already Used In Our Everyday Life*. Forbes. https://www.forbes.com/sites/bernardmarr/2019/12/16/the-10-best-examples-of-how-ai-is-already-used-in-our-everyday-life/

NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software. (2019). *NIST*. https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software

Schwartz, O. (2019, November 25). *In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation—IEEE Spectrum*. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

What is Unsupervised Learning? | IBM. (n.d.). Retrieved February 5, 2023, from https://www.ibm.com/topics/unsupervised-learning

Why is Machine Learning Important? (2021, July 6). Colorado State University Global. https://csuglobal.edu/blog/why-machine-learning-important

Wilson, E. (2022, May 31). *How to Remove Bias in Machine Learning Training Data*. Medium. https://towardsdatascience.com/how-to-remove-bias-in-machine-learning-training-data-d54967729f88