An Interactive Qualifying Project: Improving AI Technologies To Be More Inclusive

Marie Tessier

May 31, 2022

1 Abstract

Fairness in Artificial Intelligence is a growing issue in new technologies today. This paper outlines apparent social problems in Artificial Intelligence (AI) and Machine Learning (ML) and possible solutions. One solution is to provide a comprehensive guide to computer scientists and software engineers on how to integrate fairness in ML. The guide covers data collection process, pre-processing data, in-processing of the algorithm, and post-processing on how the end user is effected.

2 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are embedded in almost every industry, from healthcare applications to text, image, or speech recognition algorithms provided by companies like Amazon, Facebook, Google, IBM, and Microsoft. AI aims to mimic the problem-solving process of the human mind, while ML, a branch of AI, uses data and algorithms to imitate the way humans learn with the goal of improving accuracy.

While these technologies are useful for the healthcare industry providing diagnosis and clinical treatment [3], robotic systems [11], even in education [5], they have a larger downfall of potentially acquiring biases. As these technologies develop, biases begin to become apparent and may have extreme social impacts. One well-known example of ML is a chatbot deployed by Microsoft on Twitter, Kik, and GroupMe; this bot was developed to help in AI research by chatting with users and learn from them, but it morphed from a fun chatbot to a "neo-Nazi sexbot." The chatbot learned from many of the disturbing posts and conversations on the forum, even posting on one occasion "Bush did 9/11 and Hitler would have done a better job than the monkey we have now" [8]. Microsoft swiftly took down the chatbot after the negative Tweets that it was posting but this case is an important lesson for researchers on how quickly AI can diverge from its original intent when the data or algorithm created is not properly designed to uphold non-biased tendencies.

This technology is so interwoven in thousands of systems and programs that interact with all demographics of people, so it is critical that AI makes inclusive decisions. Without these decisions, the technologies have potential consequence of bringing further discrimination to marginalized groups. Despite this fact, many algorithms fall victim to bias due to underlying prejudices in the data itself or poor classifications within the algorithm. Bias in these technologies pose many dangers to society because it perpetuates existing bias and oppression.

Extensive literature covers topics in Inclusive AI and Fairness in ML to help combat bias within new and existing technologies. Inclusive AI is a framework on how to integrate ethical ideas on collaboration with more diverse teams and Fairness in ML refers to the attempts to correct algorithmic bias in machine learning models. Many of the documents about Inclusive AI take a more ethical, high level approach to combat bias. There are numerous writings [4, 5, 11, 12, 13] that discuss the significance of forming diverse teams of software engineers, equally representative groups, and transparency in how the algorithm makes decisions. Literature written on Fairness in ML, such as [8], primarily considers binary classification, fairness in regression, recommendation systems, unsupervised learning, and natural language. However, many of these articles still fall short on providing examples for non-Binary classification algorithms.

To address this need, the goal of this paper is to provide a comprehensive guide to computer scientists and software engineers on how to integrate fairness in ML. This goal was accomplished through the following objectives: 1) identify early signs that show a data set may have bias, 2) prevent algorithmic bias in binary classification, 3) and how the technology affects end users.

3 Background

AI aims to mimic the problem-solving process of the human mind, as Alan Turing defined AI saying; "If there is a machine behind a curtain and a human is interacting with it (by whatever means, e.g. audio or via typing etc.) and if the human feels like he/she is interacting with another human, then the machine is artificially intelligent."

Machine Learning is a branch of AI that uses data and algorithms to imitate the way humans learn and improve overall accuracy. In other words, it is a computer program that can learn behaviors that were not explicitly programmed. This behavior can be learned through the following factors: data preprocessing, data visualization and exploration, model building, and model evaluation.

Data is the backbone of how the machine learning model learns information and is used to train the algorithm. The data can be in many different formats but will always have attributes, which can be thought of as labels to a column of information. A model in machine learning is a file that has been trained to recognize certain types of patterns in a data set and can then be used to predict the behavior of future or unknown data.

While there is a lot of research on Inclusive AI, there is a gap in information on educating how to write machine learning algorithms to be fair. Without the proper regulation and education, AI can have extreme dangers in perpetuating prejudices of the past. AI is being used to make decisions in a variety of applications such as hiring, loan-approval, and parole-granting. AI can have the potential to either create a fair society or to discriminate against vulnerable groups, preventing them from being accepted for a loan or job.

To address this need, the goal of this paper is to provide a comprehensive guide to computer scientists and software engineers on how to integrate fairness in ML. This goal was accomplished through the following objectives: 1) identify early signs that show a data set may have bias, 2) prevent algorithmic bias in binary classification, 3) and how the technology affects end users.

3.0.1 Inclusive AI / Fairness in ML

The main approach to improve inclusivity within the field of AI is through a discussion within a diverse team, who are following ethical frameworks when creating the algorithm, analyzing data, and how the technology affects different user groups. This process is referred to as inclusive AI.

"For AI to be truly inclusive, changes are required at all three levels - the algorithm, the data, and the end-users. Existing ethical frameworks address mainly individual, human responsibility, not distributed responsibility" described in [9]. This relationship is seen in Figure 1, with the relationship between data, algorithm, and users need to be considered to make an inclusive AI.

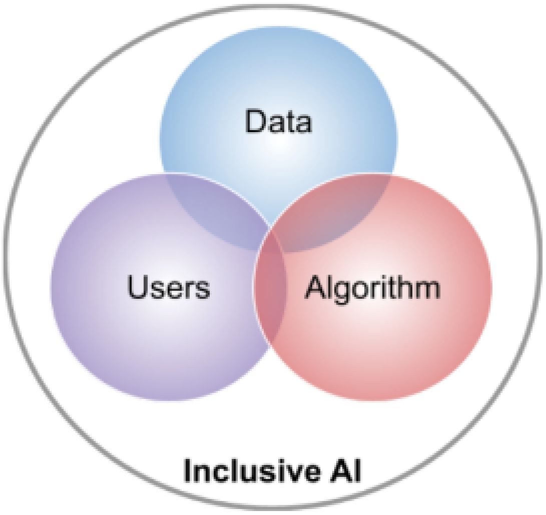


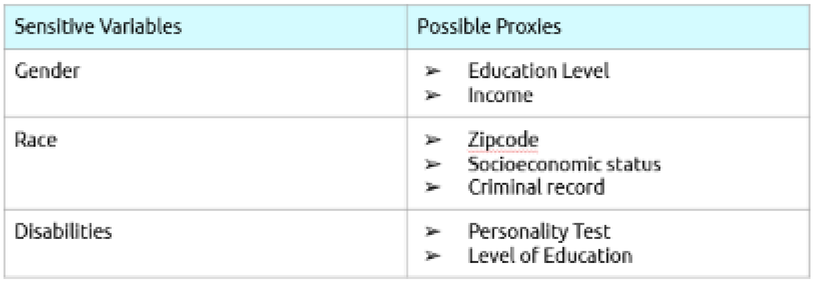
Figure 1: Parts of an Inclusive AI.

3.0.2 Data

The data source, which is the foundation to the entire process, must be scrutinized to ensure it was collected ethically and represents groups equally. When data is inappropriately used, unconscious biases can result in the technologies. I think this should be present tense: "...must be scrutinized to ensure it is collected ethically. . . "

Discuss in Fairness in machine learning: A survey [10], data needs to be collected ethically and participants must provide consent to the use of their data. There are policies such as the General Data Protection Regulation (GDPR) which is a regulation put in place by the European Union to regulate data protection and privacy.

Being able to define protected or sensitive variables for sensitive groups is critical in making sure that they are not misrepresented. Sensitive groups in the data may include, but are not limited to, race, gender, and people with disabilities [21]. Figure 2 shows possible sensitive variables with their example proxy variables.



Sensitive Variables	Possible Proxies
Gender	➢ Education Level ➢ Income
Race	➢ Zipcode ➢ Socioeconomic status ➢ Criminal record
Disabilities	➢ Personality Test ➢ Level of Education

Figure 2: Examples of sensitive variables and their proxies.

For example, big data sources that are collected through websites, social media, e-commerce vehicles, and general online searches get compiled, mined, and are eventually applied for commercial or public use. Having a large amount of data is important when training some complex ML algorithms but can also include unchecked biases. For example, in a report published by the Federal Trade Commission, an agency responsible for regulating tech companies, discuss how the misuse of big data leads to potential discriminatory harm to low-income and underserved populations [2].

The Race Disparity Unity (RDU) and Center for Data Ethics and Innovation (CDEI) partnered in 2019 to mitigate and review bias in algorithmic decision-making. In their report, they discuss the importance of understanding the data as a whole. For example, sensitive variables, such as race, may be excluded but some areas postcodes can function as a proxy variables for race. In this case, the data would still lead to a biased outcome in the algorithm [6].

Statistical parity helps to ensure that the training data set has an equal proportion of subjects. Conditional statistical parity controls for a set of plausible risk factors within an equal proportion of subjects. Predictive equality assumes that the accuracy of decisions is equal across race groups [6].

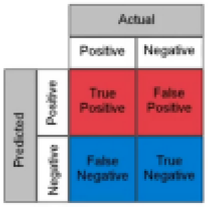
I aim to combine these techniques of sensitive and proxy variables on how to mitigate bias within the training dataset to ensure that it will be used fairly when applied to the algorithm.

3.0.3 Algorithm

In machine learning, an algorithm is a mathematical model used to map and learn patterns embedded in the data. The algorithms can perform pattern recognition, classification, and prediction on data by learning from existing data (which is known as the training set). The training set is then compared to another testing dataset to evaluate the performance of the algorithm. There are different tasks that algorithms perform, such as classification (K-nearest neighbor and support vector machine), classification/regression (linear regression, random forest and neural networks), clustering (K-means) and feature extraction (canonical correlation analysis).

There are many metrics that can be used to determine bias in a model, such as confusion matrix, group fairness, parity-based, and calibration-based metrics. Within this paper I will focus on using confusion matrix-based metrics to determine the performance and fairness of a classification model.

The confusion matrix (Figure 3) takes into consideration True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR). These metrics can be used to distinguish underlying differences between groups and regulate bias within the classifications.



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figure 3: Binary Confusion Matrix.

This measurement can be used for binary classifications to help determine the biases that may exist in the algorithm or training data.

3.0.4 User

Applications that use AI tend to have a large and diverse user base. So when looking at the end user group for a program, the developers should analyze the effects on all possible end user groups to check for bias.

For example, facial recognition software is a widely used technology for unlocking phones, tagging people on social media and even in law enforcement.

Despite the growing trend of using this technology for law enforcement, it is being banned in Boston and San Francisco for its inaccurate recognition of female, black, and 18-30 year olds. Data from the 2018 Gender Shades project revealed these discrepancies in the classification accuracy of face recognition softwares, this data is seen in Figure 4 [7].

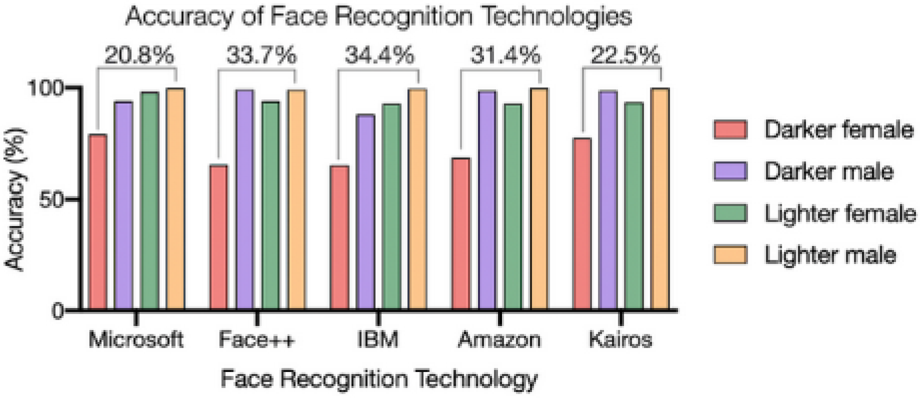


Figure 4: 2018 Gender Shades Project Accuracy of Facial Recognition.

The data from the Gender Shades project highlights that there is a 20-35% difference between the accuracy of lighter males versus darker females. The danger in this is miss identifying individuals for crimes they did not commit.

It is critical that the end user results of the ML softwares are tested on a larger user group to determine if there are any unpredicted biases toward a certain demographic.

3.1 Summary

Through this literary review I learned that there is a vast amount of information on the topics of Fairness in ML and inclusive AI. These sources have both background on the issue and possible solutions. This information can be compiled and expanded on to make a comprehensive guide to educate creators of this software to ensure a fair algorithm.

4 Methodology

The goal of this project is to create a comprehensive framework on creating unbiased AI for developers of AI and ML programs. This goal can be met though the following objectives:

1. Evaluate existing resources to identify and combat bias within the data, the algorithm and the end user's experience.

2. Compile the information learned to create a guide that goes over each section; the data, the algorithm and the end user's experience.

3. Test this process on a data set using different classification and regression algorithms. Then compare how the bias changes when this process is not used versus when it is.

5 Review and Conclusion

In this project, I worked to research the Fairness of Machine Learning and Artificial Intelligence. Where these technologies are becoming embedded into everyday life, from loan approval, law enforcement, hiring decisions and much more. it is imperative that Artificial Intelligence does not repeat prejudices of the past.

With rising problems of Machine Learning programs racial discrimination such as in 2015 Google's "photo recognition software offensively categorized two African-American people as gorillas, which was still not rectified several years later." [9] There are also examples of AI expanding the racial economic inequalities, such as found in a research paper released by a team of Berkeley researchers. These researchers found that lenders using algorithms used to determine loan pricing have discriminated against people of color, with a collective overcharge of $765 million each year for home refinance loans. Additionally, the researchers found that 1.3 million creditworthy applicants of color were rejected between 2008 and 2015 [20].

Through this project, I was able to create a guide to help developers understand and take steps to improve fairness within AI and ML models. In the future, I would like to interview professionals in the field on how helpful they find the guide. I would also like to perform my own analysis on a known bias datasets. Compare the results before using pre-processing, in-processing and post-processing fairness in ML methods outlined in this guide to see if the bias is decreased. AI technologies hold power to change the world and create a more equal opportunity for everyone.This is only if developers consider the bias that may arise and fix it.

References

[1] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-lending discrimination in the fintech era. *Journal of Financial Economics*, 143(1):30–56, 2022.

[2] the Premerger Notification Office Staff, This blog is a collaboration between CTO, DPIP staff, and the AI Strategy team. Big data: A tool for inclusion or exclusion? understanding the issues (ftc report), Aug 2021.

[3] Julia E.H. Brown and Jodi Halpern. Ai chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM - Mental Health*, 1:100017, 2021.

[4] LCS. Responsible AI and analytics for an ethical and inclusive digitized society: 20th IFIP wg 6.11 conference on e-business, e-services and e-society, I3E 2021, galway, ireland, september 1–3, 2021, proceedings. In *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society*, volume 12896 of *Lecture Notes in Computer Science*, Cham, 2021. Springer International Publishing.

[5] Haoran Xie, Gwo-Jen Hwang, and Tak-Lam Wong. Editorial note: From conventional ai to modern ai in education: Re-examining ai and analytic techniques for teaching and learning. *Educational technology & society*, 24(3):85–, 2021.

[6] Lara Macdonald. Using data to combat bias against ethnic minorities, Dec 2020.

[7] Alex Najibi. Racial discrimination in face recognition technology, Oct 2020.

[8] Rachel Metz. Why microsoft accidentally unleashed a neo-nazi sexbot, Apr 2020.

[9] Tero Avellan, Sumita Sharma, and Markku Turunen. Ai for all: Defining the what, why, and how of inclusive ai. In *Proceedings of the 23rd International Conference on Academic Mindtrek*, AcademicMindtrek '20, page 142–144, New York, NY, USA, 2020. Association for Computing Machinery.

[10] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

[11] Zachary Emil, Andrew Robbertz, Richard Valente, and Cole Winsor. Towards a more inclusive world: Enhanced augmentative and alternative communication for people with disabilities using ai and nlp. *MQP*, 2020.

[12] Jianfeng Xu, Yuanjian Zhang, and Duoqian Miao. Three-way confusion matrix for classification: A measure driven view. *Information sciences*, 507:772–794, 2020.

[13] Mark Coeckelbergh. Inclusive robotics and AI – some urgent ethical and societal issues. In *Inclusive Robotics for a Better Society*, Biosystems & Biorobotics, pages 23–24. Springer International Publishing, Cham, 2019.

[14] Paul R Daugherty, H James Wilson, and Rumman Chowdhury. Using artificial intelligence to promote diversity. *MIT Sloan Management Review*, 60(2):1, 2019.

[15] Carmen Fernández and Alberto Fernández. Inclusive ai in recruiting. multi-agent systems architecture for ethical and legal auditing. In *Highlights of Practical Applications of Survivable Agents and Multi-Agent Systems. The PAAMS Collection*, Communications in Computer and Information Science, pages 326–329. Springer International Publishing, Cham, 2019.

[16] Ameet V Joshi. Introduction to ai and ml. In *Machine Learning and Artificial Intelligence*, pages 3–7. Springer International Publishing, Cham, 2019.

[17] Niels Bantilan. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of technology in human services*, 36(1):15–30, 2018.

[18] Joyce Chou, Roger Ibars, and Oscar Murillo. In pursuit of inclusive ai. *Inclusive Design*, 2018.

[19] Emily LaRosa and David Danks. Impacts on trust of healthcare ai. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 210–215, New York, NY, USA, 2018. Association for Computing Machinery.

[20] Padala Manisha and Sujit Gujar. Fnnc: Achieving fairness through neural networks. 2018.

[21] Nicol Turner Lee. Detecting racial bias in algorithms and machine learning. *Journal of information, communication ethics in society (Online)*, 16(3):252–260, 2018.