

**Mitigating Racial Algorithmic Bias in Healthcare Artificial Intelligence Systems: A Fairness-Aware
Machine Learning Approach**

Tiffany Meg Nyawambi

96331

**Submitted in Partial fulfilment of the Requirements for the Degree of Master of Science in
Data Science and Analytics at Strathmore University**

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

2024

This dissertation is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the proposed dissertation contains no material previously published or written by another person except where due reference is made in the documentation itself.

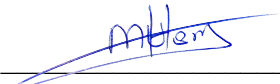
Student: Tiffany Meg Nyawambi

Sign:  Date: 05/03/2024

Approval

The dissertation proposal of Tiffany Meg Nyawambi was reviewed and approved for examination by:

Dr. Henry Muchiri,
Academic Director, Undergraduate Programmes,
School of Computing & Engineering Sciences,
Strathmore University

Sign:  Date: 07-03-2024

Abstract

The integration of machine learning in healthcare has opened new avenues for enhancing patient care and outcomes. However, these advancements come with the responsibility to ensure that algorithmic predictions, used to inform decision-making processes, do not inadvertently perpetuate, or exacerbate disparities among patient populations.

A critical examination of current research on fairness in machine learning uncovers significant gaps in both bias detection and mitigation strategies. It underscores the importance of adopting a multifaceted approach to fairness evaluation, as relying solely on one method may overlook nuanced biases and fail to address systemic issues inherent in algorithmic decision-making processes.

This paper proposes the development of classifiers using logistic regression and random forest, guided by a range of fairness metrics and bias alleviation algorithms. The fairness metrics include disparate impact, average odds difference, statistical parity difference, equal opportunity difference, and the Theil index, while the bias alleviation algorithms comprise reweighing (pre-processing algorithm), prejudice remover (in-processing algorithm), and disparate impact remover (pre-processing technique). This methodology aims to balance accuracy and fairness in healthcare machine learning applications by integrating the fairness metrics into the fundamental modeling process. It aims to employ a systematic approach, from data collection to model deployment and monitoring, to promote equitable and transparent decision-making and enhance interpretability by providing local interpretable model-agnostic explanations (LIME) for model outcomes.

Key Words: Racial Algorithmic Bias, Algorithmic Fairness, Fairness Metrics, LIME, Logistic Regression, Random Forests, Protected Classes, Interpretability

Table of Contents

Declaration.....	i
Approval	i
1. Introduction.....	1
1.1 Background to the Study	1
1.2 Problem Statement	2
1.3 Research Objectives	3
1.3.1 General Objective.....	3
1.3.2 Specific Objectives	3
1.4 Research Questions.....	3
1.5 Significance of the Study.....	3
1.6 Scope of the Study	4
1.7 Limitations of the Study.....	4
2. Literature Review	6
2.1 Introduction	6
2.2 Racial Algorithmic Bias.....	7
2.2.1 Racial Algorithmic Bias in Judicial Systems.....	7
2.2.2 Racial Algorithmic Bias in Judicial Systems.....	7
2.2.3 Racial Algorithmic Bias in Commercial artificial intelligence Classification Systems ..	8
2.2.4 Racial Algorithmic Bias in Hate Speech and Abusive Language Detection Models	8
2.3 Development of Fairness-aware Machine Learning Algorithms.....	9
2.4 Bias Mitigation Algorithms.....	11
2.4.1 Reweighing.....	11
2.4.2 Adversarial Debiasing.....	12
2.4.3 Reject Option Classification	13
2.4.4 Disparate Impact Remover	14
2.4.5 Average Odds Difference	14

2.4.6	Statistical Parity Difference.....	15
2.4.7	Theil Index	16
2.4.8	Equality of Opportunity	16
2.4.9	Prejudice Remover.....	17
2.4.10	Ethical Model Interpretability.....	18
2.5	Gaps in Existing Literature	18
2.5.1	Adversarial attacks on fairness.....	18
2.5.2	Transferability across different domains	19
2.5.3	Intersectionality.....	19
2.5.4	Unintended consequences.....	19
2.6	Conclusion	19
3.	Methodology	20
3.1	Research Design.....	20
3.2	Data Sources and Collection	21
3.3	Data Preprocessing.....	21
3.3.1	Data Cleaning	21
3.3.2	Data Exploration	21
3.3.3	Treating Missing Data.....	21
3.3.4	Treating Outliers	21
3.3.5	Data Type Conversion.....	22
3.4	Data Transformation.....	22
3.4.1	Feature Engineering.....	22
3.4.2	Feature Scaling	22
3.4.3	Feature Selection	22
3.4.4	Data Encoding	22
3.5	Model Selection.....	22
3.6	Model Evaluation.....	22

3.7	Model Deployment	23
3.8	Testing Model on Deployment Data	23
3.9	Generating Explanations for Predictions	23
3.10	Monitoring for Drift	23
3.11	Retraining the Model	23
3.12	Post-Retraining Evaluation	23
3.13	Conclusion	23
	References	24

1. Introduction

1.1 Background to the Study

Recent advancements in artificial intelligence have led to the widespread deployment of artificial intelligence systems in various domains, including finance, healthcare, and criminal justice. However, there is growing concern about the potential for algorithmic bias to perpetuate and exacerbate racial disparities in these systems. As such, it is crucial to develop fairness-aware machine learning approaches that mitigate racial algorithmic bias. Studies have highlighted the presence of biases in machine learning applications, such as facial recognition and candidate ranking, prompting active research on fairness in machine learning over the past five years (Mujtaba & Mahapatra, 2019).

Racial algorithmic bias in artificial intelligence systems occurs when the algorithms produce discriminatory outcomes that disproportionately harm, or benefit individuals based on their race. This bias can manifest in numerous ways, such as in predictive policing algorithms that disproportionately target minority communities or in healthcare algorithms that result in differential treatment based on race (Panch et al., 2018).

It is evident that algorithmic bias in artificial intelligence systems poses significant ethical and societal challenges. Researchers and practitioners have focused on developing fairness-aware machine learning approaches to address these issues, including using fairness metrics to quantify and mitigate bias in artificial intelligence systems and redesigning algorithms to ensure equitable outcomes for individuals of all races (Rajkomar et al., 2018). Furthermore, studies have highlighted the importance of incorporating fairness constraints into the machine learning process, emphasizing the need for algorithmic transparency and accountability (Fabris et al., 2023). This framework ensures that artificial intelligence systems are not only accurate and efficient but also equitable and fair in their decision-making processes, contributing to a more just and inclusive society.

In addition to technical solutions, it is crucial for policymakers and stakeholders to collaborate in establishing regulatory frameworks that govern the development and deployment of artificial

intelligence systems. Such frameworks should prioritize fairness and equity, ensuring that the benefits of artificial intelligence are distributed equitably across different racial groups and minimizing the potential for discriminatory practices (Raghavan et al., 2020).

In conclusion, addressing racial algorithmic bias in artificial intelligence systems requires a multifaceted approach that encompasses technical, ethical, and regulatory considerations. By integrating fairness-aware machine learning approaches, promoting algorithmic transparency, and enacting comprehensive regulatory frameworks, it is possible to mitigate racial bias in artificial intelligence systems and foster a more equitable and inclusive technological landscape.

1.2 Problem Statement

Racial algorithmic bias remains a persistent challenge within artificial intelligence systems, leading to differential and often unjust outcomes for individuals from marginalized racial and ethnic groups. While fairness-aware machine learning frameworks offer a potential path forward, there is a notable debate regarding their efficacy in reducing racial disparities, particularly when considering the variability of datasets and real-world scenarios (Friedman & Nissenbaum, 1996).

These approaches, while promising, have yet to undergo rigorous real-world evaluation. This lack of validation raises questions about their generalizability and the true impact they have on achieving equitable artificial intelligence systems (Raghavan et al., 2020). Moreover, the ethical deployment of artificial intelligence technologies demands that we pay close attention to the potential societal consequences, ensuring that advancements in the field do not come at the expense of social justice (Barocas & Selbst, 2016). Overcoming the issue of algorithmic bias is not solely a technical challenge but also an ethical and collaborative endeavour which requires participation from diverse stakeholders (Mujtaba & Mahapatra, 2019).

It is therefore crucial to employ and enhance existing fairness-aware frameworks to effectively combat racial algorithmic bias within artificial intelligence systems. This will involve comprehensive analytical methods, diverse dataset validation, and thoughtful consideration of ethical implications. The ultimate objective is the development of artificial intelligence systems that uphold the core principles of fairness and equity, serving all individuals equitably, regardless of their racial or ethnic background.

1.3 Research Objectives

1.3.1 General Objective

The main objective of the dissertation is to leverage existing fairness-aware artificial intelligence frameworks to effectively detect and mitigate racial algorithmic bias within healthcare artificial intelligence systems, thereby promoting equitable and transparent decision-making processes.

1.3.2 Specific Objectives

- 1) To integrate established fairness metrics into the modeling process of machine learning classifiers aimed at detecting and mitigating racial algorithmic bias within healthcare artificial intelligence systems.
- 2) To detect racial algorithmic bias in machine learning classifiers using established fairness metrics.
- 3) To mitigate identified racial algorithmic biases in the machine learning classifiers using pre-processing and in-processing bias alleviation techniques.
- 4) To assess the effectiveness of integrated fairness metrics in mitigating algorithmic bias in healthcare artificial intelligence systems.

1.4 Research Questions

- 1) How effective are different fairness metrics and algorithms in detecting and reducing bias in machine learning classifiers?
- 2) What is the impact of bias alleviation methods, such as reweighing, prejudice remover, and disparate impact remover, on reducing bias in classifiers built using Logistic Regression and Random Forest?

1.5 Significance of the Study

Algorithmic bias in machine learning classifiers poses significant challenges, particularly in areas where decisions impact individuals' lives, such as lending, hiring, and criminal justice (Rajkomar et al., 2018). Addressing racial algorithmic bias is imperative to ensure fairness, equity, and social justice. However, the effectiveness of existing fairness-aware algorithms and tools in mitigating bias needs thorough evaluation across diverse datasets and application domains.

This study seeks to fill this gap by systematically assessing the performance of these techniques and providing insights into their effectiveness and limitations. By doing so, it contributes to advancing the understanding of algorithmic bias detection and mitigation, thus facilitating the development of more equitable and transparent artificial intelligence systems.

The significance of this study lies in its potential to inform policymakers, practitioners, and researchers about the most effective strategies for detecting and mitigating racial algorithmic bias in machine learning classifiers. By systematically evaluating existing fairness metrics and algorithms, the study provides valuable insights into the practical application of these techniques across various domains. The findings of this research can guide the development and implementation of more equitable and transparent artificial intelligence systems, thereby reducing the potential for discriminatory outcomes and promoting fairness and social justice. Additionally, the study contributes to building trust in artificial intelligence technologies by demonstrating their ability to uphold principles of fairness and equity, ultimately benefiting society.

1.6 Scope of the Study

The study primarily focuses on evaluating the effectiveness of existing fairness metrics and algorithms for detecting and mitigating bias within machine learning classifiers. It includes an examination of bias alleviation methods, such as reweighing, prejudice remover, and disparate impact remover, applied to classifiers constructed using Logistic Regression and Random Forests. Additionally, the research encompasses an analysis of the explanations generated for model predictions using LIME. The study aims to provide insights into addressing racial algorithmic bias within artificial intelligence systems by leveraging existing tools and methodologies.

1.7 Limitations of the Study

Firstly, the effectiveness of bias detection and mitigation techniques may vary depending on the specific dataset and application domain, which may limit the generalizability of the findings. Secondly, the study relies on existing fairness-aware algorithms and tools, which may not cover all possible scenarios or capture every aspect of algorithmic bias. Furthermore, the analysis may be constrained by the availability and quality of data, potentially affecting the robustness of the

results. Additionally, the study does not explore novel solutions or develop new algorithms but rather focuses on the application and evaluation of existing methodologies.

2. Literature Review

2.1 Introduction

In today's rapidly evolving world, artificial intelligence systems play a critical role in various domains, including talent assessment and recruitment processes (Johndrow & Lum, 2019). These systems are utilized by companies and governments to inform hiring decisions, employee management, policing, credit scoring, insurance pricing, and many other aspects of our society. However, the increasing reliance on artificial intelligence systems has raised concerns about the potential for algorithmic bias, especially racial bias, in decision-making processes.

Current research on algorithmic fairness underlines the complex nature of bias in artificial intelligence systems. Studies have identified biases in various machine learning applications like facial recognition and candidate ranking, pointing to the critical need for fairness in machine learning (Mujtaba & Mahapatra, 2019). Other recent investigations demonstrate how even sophisticated algorithms can produce discriminatory outcomes; for example, some algorithms recommend pretrial release of white defendants at higher rates than Black defendants with identical risk profiles (Arnold et al., 2021).

This literature review aims to comprehensively examine existing research efforts aimed at identifying instances of racial bias within artificial intelligence systems across various domains. By synthesizing and analyzing the findings from diverse studies, the review aims to gain a deeper understanding of the nuanced manifestations of algorithmic bias and its implications for decision-making processes. Additionally, the review aims to explore the efficacy of different strategies and methodologies proposed for mitigating racial algorithmic bias. This includes examining the development and implementation of fairness-aware machine learning algorithms, as well as investigating the ethical and societal implications of biased artificial intelligence systems. The review seeks to identify promising avenues for addressing algorithmic bias and promoting fairness and equity within artificial intelligence systems by critically evaluating existing approaches and methodologies.

Furthermore, this literature review endeavours to identify gaps and limitations in current research efforts, to pave the way for future investigations in this area. This will contribute to the

advancement of knowledge and the development of more robust and equitable artificial intelligence systems. Overall, the objectives of this literature review are to provide a comprehensive overview of existing research on mitigating racial algorithmic bias and to identify opportunities for future research and intervention.

2.2 Racial Algorithmic Bias

2.2.1 Racial Algorithmic Bias in Judicial Systems

Studies by (Arnold et al., 2021) have made empirical findings about machine learning algorithms in the judicial system, where there is evidence that racial discrimination persists even when race and ethnicity are not explicitly included in the training data. The use of sophisticated machine learning algorithms and regression-based recommendations in pretrial risk assessments discriminates against Black defendants, showing a significant unwarranted racial disparity in algorithmic recommendations. This unwarranted disparity in trial release rates between white and Black defendants with identical risk profiles accounts for a sizable portion of the observed racial disparities in algorithm recommendations. By extrapolating quasi-experimental variation across decision-makers, it has been possible to estimate moments of algorithmic discrimination in New York City. These studies shed light on the presence of racial algorithmic bias in machine learning algorithms used in the criminal justice system.

2.2.2 Racial Algorithmic Bias in Judicial Systems

(Sweeney, 2013) Investigated the existence of racial bias in online ad delivery systems by analyzing whether ads for high-paying jobs were shown more frequently to users with "white-sounding" names compared to those with "black-sounding" names. Using a controlled experiment, they found evidence of racial discrimination, with job ads for high-paying positions being shown more frequently to users with "white-sounding" names indicating that the algorithms used by these ad delivery systems perpetuated and amplified existing racial biases in employment opportunities.

(Lee, 2018) Highlights the need for research to understand how emerging technologies, such as artificial intelligence systems, can perpetuate unfair and disparate treatment for certain populations and stress the importance of addressing algorithmic bias to ensure fairness and equity in employment processes.

2.2.3 Racial Algorithmic Bias in Commercial Artificial Intelligence Classification Systems

In their seminal work, (Buolamwini & Gebru, 2018) take a critical look at the accuracy of commercial artificial intelligence systems in classifying gender, especially examining how these systems perform across different skin tones and gender presentations. In their investigation, they assessed several commercial facial analysis tools and discovered that the systems had higher error rates for women, particularly for darker-skinned women. The results showed that lighter-skinned males were the most accurately recognized group, while darker-skinned females were the least accurately recognized. This disparity in performance is highly indicative of racial bias embedded within the machine learning models. The training datasets used for these systems often have a disproportionately high number of lighter-skinned subjects. As a result, the algorithms are less accurate when identifying individuals who do not fit these demographic categories. Addressing these concerns, the researchers call for more inclusive and diverse datasets to train artificial intelligence models and for greater transparency and accountability from companies developing these technologies.

2.2.4 Racial Algorithmic Bias in Hate Speech and Abusive Language Detection Models

(Sap et al., 2019) tackle racial bias from a different angle, looking into hate speech and abusive language detection models. The research team analyzed various datasets typically used to train algorithms for detecting hate speech and abusive language online. Their investigation centered on understanding how the labels and annotations assigned to text samples could reflect racial biases. Annotations in these datasets classify texts as hate speech, abusive, or neither; but this binary oversimplification often fails to capture nuanced expressions of bias. Upon analyzing the labels and the context of annotated examples, the researchers found that the language associated with certain racial groups was more likely to be deemed hate speech or abusive, regardless of the intent or content. This indicates a systemic issue where the labelling process itself is subject to the annotators' biases, which may stem from societal stereotypes and prejudices. Such biases in training datasets can have profound implications when the resultant models are deployed in the real world. They can lead to disproportionate flagging or censorship of content from specific racial groups, infringing their freedom of expression and access to online platforms. This study

contributes significantly to ongoing efforts to address algorithmic discrimination by highlighting the complex challenges in designing fair and unbiased artificial intelligence systems. It underscores the need for continuous re-evaluation and enhancement of datasets, algorithms, and practices to move towards truly fair and unbiased artificial intelligence mechanisms.

2.3 Development of Fairness-aware Machine Learning Algorithms

The development of fairness-aware machine learning algorithms is a crucial step in mitigating racial algorithmic bias in artificial intelligence systems. These algorithms aim to address and reduce the disparities and discriminatory outcomes that arise from biased data and biased decision-making processes (Friedler et al., 2019). Fairness-aware machine learning algorithms consider fairness criteria during the training and decision-making process. They incorporate techniques such as reweighing of input samples and regularization to counteract the impact of biased training data (Du et al., 2021). Moreover, these algorithms strive to ensure that the predictions and outcomes of artificial intelligence systems are not influenced by protected attributes such as race or ethnicity. Fairness-aware machine learning algorithms also recognize the importance of transparent and explainable artificial intelligence systems. They aim to provide clear and interpretable insights into how decisions are made, allowing for better understanding and accountability (Lee, 2018).

A study by (Obermeyer & Mullainathan, 2019) dives deep into the complexities emblematic of healthcare data and decision-making processes, which are often entangled with socioeconomic, racial, and gender biases. These biases can inadvertently be encoded into predictive algorithms when the training data reflect historical disparities or when the design of these systems does not adequately account for diverse patient populations. They emphasize that merely adjusting statistical or performance measures of an algorithm is not enough to ensure fairness and assert that a fairness-aware approach must account for the diverse ways in which bias can manifest in healthcare algorithms. This includes considering outcome equity across diverse groups, ensuring that performance, like predictive accuracy, does not disproportionately favour one group over another, and looking at equitable resource distribution guided by the needs and benefits to various populations. To mitigate these concerns, they suggest several methodologies that can be used to make machine learning models in healthcare more fairness aware. These methods may include

techniques like recalibration of algorithms across groups, the incorporation of fairness constraints during the model training process, and the design of novel algorithms that prioritize equity as a performance metric. They further call for ongoing evaluation and testing of healthcare algorithms in real-world settings to continually assess and refine their fairness. The inclusion of healthcare providers, patients, and other stakeholders in the evaluation process is crucial for the nuanced understanding of how these models operate in practice.

(Calmon et al., 2017) provides valuable insights into operationalizing fairness in machine learning applications. The authors highlight key areas such as the multiplicity of fairness definitions, the selection of appropriate algorithmic techniques for fairness interventions, and methodologies for evaluating fairness within machine learning models. A significant aspect of this research is its pragmatic approach. It recognizes that fairness is context-dependent, and a one-size-fits-all definition is often inadequate. The paper contributes to the understanding that different applications may require different fairness criteria, and it is crucial to align the choice of fairness definition with the specific social and ethical objectives of the application at hand. Another critical highlight is the discussion on the complexity of translating abstract fairness definitions into concrete algorithmic forms. The study explores various techniques, such as pre-processing, in-processing, and post-processing, which can be implemented at various stages of algorithm development to mitigate bias. Finally, the paper underscores the importance of rigorously evaluating the effectiveness of fairness interventions through various metrics and validation methods. It also addresses the challenges of operationalizing fairness in real-world scenarios, such as the trade-offs between fairness and model performance or the difficulty in obtaining representative data. This study stands out for its focus on the practical implementation of fairness-aware machine learning and contributes to the field by sharing experiences and insights gleaned from real-world attempts to create more equitable algorithms. By laying out a comprehensive framework for analyzing and applying fairness-aware techniques, the survey provides valuable guidance for both researchers and practitioners looking to navigate the complex terrain of fairness in artificial intelligence. It serves as a call to action for continued innovation and interdisciplinary collaboration in the journey toward more equitable machine learning systems.

(Barocas et al., 2017) presents a significant argument that fairness in sociotechnical systems transcends technical fixes and demands an understanding of these systems' social contexts. They

contend that fairness should be conceived with a broader perspective that reflects societal complexities, power dynamics, and equity considerations. This concept recognizes the impact of technology on different social groups, emphasizing the need to protect historically marginalized communities. The authors call for a multidisciplinary approach to develop equitable sociotechnical systems and stress the importance of involving various stakeholders, including the public, in guiding technology development and governance. Their work sets a comprehensive agenda for future research in creating fairness-aware algorithms that contribute to a more inclusive society.

(Pleiss et al., 2017) Investigate whether enhancing fairness through adjustments to the algorithm might compromise the model's calibration, and vice versa, highlighting the potential balance or tension between these two objectives. The study employs both theoretical and empirical methods to evaluate how fairness interventions impact calibration. Fairness is typically assessed using metrics that determine if all groups are treated equitably, such as demographic parity or equalized odds. Calibration, on the other hand, involves metrics like expected calibration error, which measures how closely a model's predicted probabilities match the actual outcome frequencies. The findings of the study provide insights into the trade-offs between fairness and calibration that practitioners must navigate. The research demonstrates that interventions for bias mitigation can indeed influence the calibration of a model, which holds significant implications for the deployment of fair machine learning systems in real-world settings. By offering practical recommendations, the study aids in the evaluation and understanding of these trade-offs, suggesting ways to achieve a balanced outcome. This is particularly relevant for fields where both fairness and accurate probability assessments are critical, such as criminal justice risk assessments or credit scoring.

2.4 Bias Mitigation Algorithms

2.4.1 Reweighting

Reweighting is a data preprocessing technique applied to reduce bias in machine learning models. It functions by assigning different weights to the instances in the training dataset to correct imbalances related to protected attributes, such as gender or race. The goal is to neutralize any unfair advantage or disadvantage a model might learn from the frequency of certain groups within the data, ensuring more equitable treatment of different demographic groups (Celis et al., 2020).

A key study that incorporates reweighing is by (Kamiran & Calders, 2011), where they introduce the concept within the context of classification without discrimination. Their approach was named 'Reweighing,' as it assigns weights to data objects to make the dataset discrimination-free. They propose two solutions: creating weighted tuples or resampling the dataset so that the discrimination is removed. The authors' empirical evaluation showed that the reweighing technique could significantly reduce discrimination in the predictive models without degrading their performance. Specifically, by employing their reweighing process, they saw enhancements in discrimination-aware classification metrics, suggesting that their proposed fairness interventions could achieve near-equal false positive and false negative rates across different demographic slices of data.

In conclusion, reweighing serves as an effective and relatively straightforward method for bias mitigation within the preprocessing phase of model development. This strategy, particularly as discussed in their study, supports the development of fairer machine learning models, maintaining the balance between model utility and fairness. As artificial intelligence continues to drive critical decisions in various societal domains, the role of techniques like reweighing becomes ever more paramount, ensuring that advancements in technology foster inclusivity and ethical fairness.

2.4.2 Adversarial Debiasing

Adversarial debiasing denotes a unique approach within the machine learning domain aimed at countering bias through an adversarial training framework. This technique involves a dual-model dynamic where the primary model endeavors to perform the primary task effectively, while an adversarial model aims to predict the sensitive attribute, such as gender or ethnicity, using the primary model's predictions. Through this adversarial process, the main model is encouraged to learn representations that do not encode information about the protected attributes, thus promoting fairness (Zhang et al., 2018).

One notable study that examines adversarial debiasing is by (Zhang et al., 2018). In this research, the authors apply adversarial debiasing to the context of word embeddings, which are critical components in natural language processing. They construct a framework through adversarial learning that compels the development of less biased data representations. Their approach demonstrated that adversarial debiasing could successfully mitigate undesired biases present in word embeddings by preventing the adversarial network from determining the protected attribute

based on the embeddings generated by the main model. The findings from their study indicate that adversarial debiasing is capable of significantly reducing bias in word embeddings, thereby addressing sexist and ethnocentric prejudices inherent in the original data. This is evidenced by the adversarial model's diminished capability to predict the gender attribute from the debiased embeddings.

Such studies bolster the credibility of adversarial debiasing as a potent mechanism for integrating fairness more intrinsically within machine learning models. Unlike techniques that purely focus on altering the input data or adjusting outputs post hoc, adversarial debiasing embeds the fairness criterion directly into the learning algorithm, offering a path toward developing artificial intelligence systems that inherently respect the tenets of fairness and anti-discrimination.

2.4.3 Reject Option Classification

Reject option classification is a methodology in machine learning that addresses bias by introducing a decision postponement for instances near the classification boundary, where the probability of making potentially biased predictions is higher. By implementing a reject option, the classifier is allowed to withhold making a definitive prediction when it lacks confidence, thereby reducing the chance of biased outcomes based on ambiguous information (Brinkrolf & Hammer, 2018).

A notable study by (Pedreshi et al., 2008) explores this concept in their work. They provide an analysis of reject option classification as a discrimination-aware technique, focusing on how it affects the decision-making process when uncertainty is present. Their approach involves defining a decision boundary and a corresponding reject option zone where the model avoids making decisions that could be influenced by biased or uncertain data. The study found that introducing a reject option where decisions are deferred due to high uncertainty helps to alleviate bias in the predictions. This is particularly true for instances bordering the decision boundary, which are typically the most vulnerable to discrimination. Their findings highlight the effectiveness of reject option classification in promoting fairness when dealing with sensitive attributes in datasets.

Reject option classification, as shown, offers a principled way to incorporate considerations of equity into predictive models. By accepting that some decisions cannot be made fairly without

further investigation, this approach advocates for a more cautious and ethical application of machine learning, especially in high-stakes situations where biased decisions can have profound implications.

2.4.4 Disparate Impact Remover

Disparate Impact Remover is a data pre-processing technique aimed at reducing bias by adjusting feature values to obscure the correlation between attributes and the protected characteristics, like race or gender, without altering the predictive distributions of each group. This method attempts to balance the representation across different groups, thus mitigating potential disparate impacts, which occur when policies or practices have a disproportionately adverse effect on a protected group, even if unintended.

(Feldman et al., 2015) present a thorough exploration of this approach, where they discuss the concept of the Disparate Impact Remover as a technique for achieving demographic parity in predictions. By adjusting the distributions of features to be more uniform across different groups, their methodology aims to ensure that decision processes are free from biases that could lead to disparate impact. Their research demonstrates that using the Disparate Impact Remover can be effective in reducing discrimination. They found that altering feature values to diminish the correlation with protected attributes can lead to decisions that are less likely to negatively affect any group based on those attributes. This is of particular importance in scenarios such as hiring, lending, and admissions, where decisions can have a substantial impact on individuals' lives.

By contributing a pragmatic solution to the problem of bias in automated decision-making, their study reveals the potential of data pre-processing techniques like the Disparate Impact Remover in fostering fairer machine learning applications. It underscores the capacity for algorithmically driven systems to incorporate considerations of social equity, advancing the pursuit of fairness in artificial intelligence and machine learning.

2.4.5 Average Odds Difference

(Pleiss et al., 2017) define Average Odds Difference as a fairness metric used in machine learning to measure discrepancies across different demographic groups by comparing the difference in false positive rates and true positive rates. The researchers analyze how enforcing certain fairness

constraints can affect the calibration of a machine learning model and propose calibration methods that can achieve both fairness and accuracy. The authors found that minimizing average odds difference helps to ensure fair treatment of individuals across groups defined by sensitive attributes. Their empirical analysis presents a nuanced understanding of the tension between calibration and fairness, particularly emphasizing situations where achieving both may be challenging. The study offers methodological insights into how fairness interventions such as adjusting average odds difference can influence the overall balance of a predictive model's performance, with implications for deploying fair artificial intelligence systems in practice.

Their work adds to the growing body of literature on fairness in machine learning by elaborating on the practical trade-offs and considerations when striving for equitable machine learning outcomes. By focusing on average odds difference, the authors contribute to the critical discussions on how to develop algorithms that make unbiased decisions and serve all demographics justly.

2.4.6 Statistical Parity Difference

Statistical Parity Difference is a fairness criterion in machine learning that measures the difference in the probability of favourable outcomes for different demographic groups defined by a protected attribute such as race or gender (Zhang, 2022).

(Feldman et al., 2015) explore how standard supervised learning algorithms can inadvertently lead to discrimination even in the absence of explicit demographic information. They propose a methodology to adjust the data to remove disparate impact as per statistical parity, essentially certifying that the predictions do not suffer from unlawful discrimination according to this criterion. The findings of their research suggest that it is possible to modify a classifier to satisfy statistical parity without substantially reducing the accuracy of the model. This study provides a valuable perspective on how one can intervene in a machine learning pipeline to address potential biases that might result in unfair treatment of individuals based on protected characteristics.

By investigating statistical parity difference with a focus on both fairness and legal implications, the work of these researchers adds important insights into the development of machine learning systems that respect social and legal standards of equality.

2.4.7 Theil Index

The Theil Index is a measure adapted from economics to assess inequality and has been applied in machine learning to evaluate fairness across different demographic groups. A relevant study is by (Speicher et al., 2018) which discusses different fairness measures, including the Theil Index, and compares their utility in assessing equity within artificial intelligence systems. The researchers examine the application of the Theil Index to quantify inequality in the distribution of outcomes produced by machine learning models. They investigate how this economic index can be repurposed to study fairness in predictions and shed light on how inequality in such predictive outcomes can reflect and reinforce existing social disparities. The study finds that the Theil Index is a valuable tool for capturing both group-level and individual-level fairness. The index's decomposition properties allow it to identify between-group and within-group inequality, making it a flexible and informative measure for diagnosing and addressing unfairness in machine learning predictions.

The research advances the discourse on ethical artificial intelligence by demonstrating the significance of economic inequality measures for developing algorithms that are just and equitable across various populations by incorporating the Theil Index into machine learning fairness analysis. This approach provides a granular understanding of how fairness can be quantitatively assessed and improved in artificial intelligence systems, promoting the responsible use of technology in society.

2.4.8 Equality of Opportunity

Equality of Opportunity is a principle centered on the idea that all individuals should have an equal chance to achieve certain outcomes, without being hindered by arbitrary factors such as race, gender, or socioeconomic status. Within the context of fair machine learning and decision-making systems, it seeks to ensure that all individuals have comparable chances of success when irrelevant factors are disregarded, emphasizing a fair and meritocratic process (Elzayn et al., 2018), (Carey & Wu, 2022), (Rouse & Kemple, 2009).

(Hardt et al., 2016) propose a framework that centers around the principle of equality of opportunity, which requires that individuals who are similar concerning a task-relevant outcome should receive similar predictions, regardless of their membership in any demographic group. This strategy is aimed at mitigating bias by ensuring that all groups have an equal chance of receiving

beneficial outcomes when they are equally qualified. To assess the fairness of algorithms, the framework employs a metric based on the notion of equal opportunity and focuses on reducing disparities in error rates across different groups. The evaluation involves both theoretical analysis and empirical validation, where fairness-aware algorithms are assessed based on their ability to achieve equalized odds and equal opportunity in predictions. The findings from their research indicate that the proposed framework can effectively identify and mitigate unfair disparities in machine learning algorithms. The paper proves both theoretically and empirically that applying the concept of equality of opportunity allows for improved fairness in the outcomes of supervised learning systems. By equalizing the opportunity across demographic groups, the study's framework represents an influential contribution to the development of bias mitigation strategies in artificial intelligence, promoting a more equitable treatment of individuals. Through their conceptual and methodological innovations, the authors of this study provide a significant step forward in the pursuit of fairness in artificial intelligence. Their framework sets a standard for what it means to achieve equality of opportunity in machine learning and offers a clear benchmark for future research and practice in the field of algorithmic fairness.

2.4.9 Prejudice Remover

The Prejudice Remover is an in-processing method for mitigating bias in machine learning models by adding a regularization term that discourages prejudicial associations in the data. This technique goes beyond pre- and post-processing methods by adjusting the learning algorithm itself to prioritize fairness alongside accuracy.

In their research of the Prejudice Remover algorithm, (Kobayashi et al., 2012) propose an approach that incorporates a prejudice index into the objective function of a learning algorithm, which encourages the model to minimize prejudice along with error. The study demonstrates that by using the Prejudice Remover, machine learning models can reduce bias in decisions without seriously compromising their predictive accuracy. The researchers provide empirical evaluations to show how the prejudice remover can be effectively applied across various data sets and tasks, substantiating its utility and effectiveness in creating fairer artificial intelligence systems.

This work sheds light on the potential for machine learning to be engineered with an intrinsic sensitivity to fairness. It signals a shift towards an initiative-taking stance in algorithm

development where fairness is addressed directly in the model-building phase, potentially leading to more ethically aligned technology solutions.

2.4.10 Ethical Model Interpretability

Rather than presenting a specific bias mitigation strategy, (Lipton, 2018) challenges the notion that interpretability is an unalloyed good. In his paper, he discusses how the quest for interpretable models might, in some cases, lead to unexpected ethical pitfalls, such as perpetuating existing biases or introducing new forms of discrimination. The evaluation does not use typical fairness metrics; instead, it offers a philosophical and conceptual critique of interpretability in machine learning. He discusses the multifaceted nature of interpretability and suggests that it should not be pursued at the expense of other important considerations like fairness.

This study contends that the appeal of interpretability can sometimes overshadow critical awareness of how models might perpetuate bias. It emphasizes the importance of a critical examination of how efforts to make models interpretable may interact with, or even counteract, efforts to make them fair. The paper highlights the trade-offs between interpretability and fairness and calls for the machine learning community to recognize and navigate these trade-offs thoughtfully. He proposes a more nuanced approach to developing artificial intelligence systems—one that considers interpretability as one factor among many, rather than an end, in the pursuit of ethical machine learning.

2.5 Gaps in Existing Literature

There is a growing interest in algorithmic fairness and the development of methods to mitigate bias in machine learning models (Beutel et al., 2019). However, the following gaps are present in the existing literature when it comes to understanding how fairness can be directly addressed in the model-building phase.

2.5.1 Adversarial attacks on fairness

While the algorithms focus on ensuring fairness under nominal conditions, there may be insufficient research on how these models perform under adversarial attacks specifically aimed at exploiting fairness constraints. Studies often evaluate fairness at a single point in time, whereas the long-term impact of enforcing fairness through these metrics and methods remains under-explored.

2.5.2 Transferability across different domains

Fairness-enhancing interventions may not be universally effective across various domains and contexts. Determining the conditions under which these algorithms work best is essential for practical applications. Bias in machine learning models has significant real-world consequences, particularly for marginalized communities (Friedler et al., 2018).

2.5.3 Intersectionality

Many fairness metrics, including those mentioned, typically focus on individual sensitive attributes. There is a need for more research on how to best manage intersectionality, where individuals belong to multiple protected groups simultaneously. Ensuring that the fairness metrics themselves are valid and reliable indicators of bias is critical.

2.5.4 Unintended consequences

There is a possibility that interventions which seek to optimize for a particular definition of fairness could lead to unintended negative consequences, such as a decrease in model utility for non-protected groups.

Techniques that aim to improve fairness might require access to sensitive attributes, which could raise privacy concerns that need to be balanced effectively.

2.6 Conclusion

In conclusion, this literature review has presented a critical examination of the current research landscape surrounding racial algorithmic bias within artificial intelligence systems. The gaps identified underscore the challenge of developing fairness-aware machine learning methodologies capable of addressing complex, multifaceted issues of bias and discrimination.

The literature review has laid a foundation for meaningful progress in the field of artificial intelligence fairness. It challenges researchers, technologists, and policymakers to work collaboratively on harnessing the power of machine learning to foster equity and justice. As the complexities of artificial intelligence continue to have impact on society, it becomes increasingly crucial to ensure that progress in technology equates to progress in social values, particularly in the pursuit of racial fairness. Thus, this chapter stands as a call to action, urging the field to move

beyond mere acknowledgment of biases, toward active implementation of solutions that address them comprehensively.

3. Methodology

3.1 Research Design

The advent of machine learning in healthcare promises transformative changes in the diagnosis, treatment, and management of diseases. As these technologies advance, they hold the potential to improve patient outcomes, streamline care processes, and optimize resource allocation. However, the benefits of these innovations can only be fully realized when the models that underlie them are both effective and equitable. This necessitates a methodology that is diligent about not only achieving high accuracy but also ensuring fairness in predictions. It is within this context that the methodology for this study has been meticulously crafted.

Healthcare, by its very nature, deals with diverse populations, each with unique characteristics and needs. This diversity and complexity necessitate the use of sophisticated algorithms capable of capturing nuanced patterns within vast datasets. Equally important is the commitment to fairness, as decisions informed by these algorithms can have significant implications for individuals' well-being. The proposed methodology comprehensively aims to address these dual priorities by incorporating fairness metrics into the core of modeling processes. The methodology will follow a systematic, iterative approach—from data collection to model deployment and monitoring—designed to scrutinize and mitigate bias, thus will ensure that the models are just and the care recommendations they provide are equitable.

The focus is on creating models that will not only perform with high accuracy but also respect ethical guidelines and comply with legal standards regarding the treatment of protected classes. This study will be conducted to systematically detect and attenuate bias in machine learning models in the healthcare sector. The research will involve building classifiers using logistic regression and random forests, guided by a selection of fairness metrics such as disparate impact, average odds difference, statistical parity difference, equal opportunity difference, and the Theil index. The study's dual objectives are to assess and improve fairness in algorithmic predictions and provide interpretability through LIME-generated explanations for model outcomes.

3.2 Data Sources and Collection

The data will be collected from the Medical Expenditure Panel Survey. The dataset will include a diverse representation of patients, spanning different racial and ethnic backgrounds, to ensure comprehensive analysis and evaluation of algorithmic fairness. A robust data governance process ensures that data usage complies with all relevant legal and ethical guidelines, particularly around protected classes. The dataset will be partitioned into learning (50%), validation (30%), and testing (20%) sets to facilitate a comprehensive assessment of the models.

3.3 Data Preprocessing

3.3.1 Data Cleaning

An extensive initial audit of the dataset will be performed to identify and correct errors, discrepancies, and irrelevant data entries that could corrupt the model's learning process. This step will ensure a high level of data integrity and accuracy.

3.3.2 Data Exploration

Descriptive statistics and visualization tools will be utilized to explore underlying patterns and distributions, exposing the data's structure, and highlighting any preliminary biases or anomalies in the dataset.

3.3.3 Treating Missing Data

The patterns and mechanisms causing data to be missing will be evaluated and imputation techniques applied to manage the missing data ensuring the completeness of the dataset without introducing substantial bias.

3.3.4 Treating Outliers

Robust statistical methods will be employed to detect and handle outliers, ensuring they are appropriately included or excluded based on their potential impact on the overall predictive model performance and fairness. The treatment of outliers will consider the healthcare context to ensure meaningful data points are not inadvertently removed.

3.3.5 Data Type Conversion

All categorical data will be examined and converted into numerical values through encoding techniques. This process will be critical for ensuring compatibility with machine learning algorithms and for aiding in the interpretive analysis.

3.4 Data Transformation

3.4.1 Feature Engineering

This critical step involves creating new features informed by domain knowledge to improve model predictability and fairness. Feature engineering will be particularly focused on capturing nuances in healthcare utilization that affect fairness relative to protected classes.

3.4.2 Feature Scaling

Numeric features will undergo scaling to ensure that the range of the data does not unduly influence the algorithm, which can disproportionately affect the weight of features in predictive models.

3.4.3 Feature Selection

Using a combination of statistical techniques and domain expertise, the most predictive features that contribute to a fair and accurate model will be identified and selected.

3.4.4 Data Encoding

Categorical variables will undergo encoding methods suitable for the models in question. Care will be taken to choose encoding methods that support fairness in the model and avoid introducing additional bias.

3.5 Model Selection

With fairness as a pivotal goal, logistic regression and random forest classifiers will be selected due to their interpretability and effectiveness, respectively. Each model's feasibility for fair predictions will be critically assessed.

3.6 Model Evaluation

The models will be evaluated based on performance metrics including accuracy, as well as fairness tests applying metrics such as disparate impact and Theil index. The evaluation will ensure adherence to fairness as dictated by relevant regulations.

3.7 Model Deployment

The finalized models will be deployed in an enterprise application designated for care management prioritization. This application will utilize the models to score healthcare utilization, producing results and recommendations that will then be provided to care management professionals, such as nurses. The deployment process will include:

3.8 Testing Model on Deployment Data

The models will be tested on real-world deployment data to ensure that their predictions remain consistent with the training and validation phases.

3.9 Generating Explanations for Predictions

Using LIME, the study will generate interpretable explanations for each recommendation made by the model, ensuring that care management professionals can understand and trust the model's output.

3.10 Monitoring for Drift

Continual evaluation of the model's predictions will occur to detect any drift from the model specifications outlined in its factsheet. Significant deviations will trigger reevaluation and potential retraining of the model.

3.11 Retraining the Model

When drift is observed, or feedback indicates a necessary recalibration, the model will be sent back for retraining with updated data, and possibly new features or algorithm adjustments to maintain performance and fairness standards.

3.12 Post-Retraining Evaluation

Once retrained, the model will be tested again on deployment data to ensure that the adjustments have effectively addressed any identified issues without introducing new ones.

3.13 Conclusion

The methodology closes with a commitment to ethical artificial intelligence deployment, highlighting that the primary goal of implementing these machine learning models is to support equitable and just healthcare outcomes. The process outlined in this methodology ensures rigorous

development, evaluation, and ongoing maintenance of models so they may serve as reliable and unbiased tools in the healthcare industry, ultimately contributing to improved patient care and outcomes for all individuals, regardless of their racial or ethnic background.

References

1. Arnold, D H., Dobbie, W., & Hull, P. (2021, May 1). Measuring Racial Discrimination in Algorithms. AEA papers and proceedings, 111, 49-54.
<https://doi.org/10.1257/pandp.20211080>
2. Barocas, S., & Selbst, A D. (2016, January 1). Big Data's Disparate Impact. RELX Group (Netherlands). <https://doi.org/10.2139/ssrn.2477899>
3. Barocas, S., Hardt, M., & Narayanan, A. (2017, January 1). Fairness and machine learning. <https://fairmlbook.org/>
4. Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J., & H., E. (2019, January 27). Putting Fairness Principles into Practice.
<https://doi.org/10.1145/3306618.3314234>
5. Brinkrolf, J., & Hammer, B. (2018, April 1). Interpretable machine learning with reject option. *Automatisierungstechnik*, 66(4), 283-290. <https://doi.org/10.1515/auto-2017-0123>
6. Buolamwini, J., & Gebru, T. (2018, January 21). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. , 77-91.
<https://www.media.mit.edu/publications/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification/>
7. Calmon, F P., Wei, D., Vinzamuri, B., Ramamurthy, K N., & Varshney, K R. (2017, January 1). Optimized Pre-Processing for Discrimination Prevention. *neural information processing systems*, 30, 3992-4001.
https://krvarshney.github.io/pubs/CalmonWVRV_nips2017.pdf
8. Carey, A N., & Wu, X. (2022, June 25). The statistical fairness field guide: perspectives from social and formal sciences. *Springer Nature*, 3(1), 1-23.
<https://doi.org/10.1007/s43681-022-00183-3>

9. Celis, E., Keswani, V., & Vishnoi, N K. (2020, July 12). Data preprocessing to mitigate bias: A maximum entropy based approach. , 1, 1349-1359.
<https://www.semanticscholar.org/paper/Data-preprocessing-to-mitigate-bias%3A-A-maximum-Celis-Keswani/8606a6cd844de12eb8690751d25b3d019e18a55d>
10. Du, M., Yang, F., Zou, N., & Hu, X. (2021, July 1). Fairness in Deep Learning: A Computational Perspective. *IEEE Intelligent Systems*, 36(4), 25-34.
<https://doi.org/10.1109/mis.2020.3000681>
11. Elzayn, H., Jabbari, S., Jung, C., Kearns, M., Neel, S., Roth, A., & Schutzman, Z. (2018, August 30). Fair Algorithms for Learning in Allocation Problems. Cornell University.
<https://doi.org/10.48550/arxiv.1808.10549>
12. Fabris, A., Baranowska, N., Dennis, M., Hacker, P., Saldivar, J., Borgesius, F Z., & Biega, A J. (2023, September 25). Fairness and Bias in Algorithmic Hiring. Cornell University. <https://doi.org/10.48550/arxiv.2309.13933>
13. Feldman, M B., Friedler, S A., Moeller, J F., Scheidegger, C., & Venkatasubramanian, S. (2015, August 10). Certifying and Removing Disparate Impact.
<https://doi.org/10.1145/2783258.2783311>
14. Friedler, S A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E P., & Roth, D. (2018, February 12). A comparative study of fairness-enhancing interventions in machine learning. <https://arxiv.org/abs/1802.04422>
15. Friedler, S A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E P., & Roth, D. (2019, January 29). A comparative study of fairness-enhancing interventions in machine learning. <https://doi.org/10.1145/3287560.3287589>
16. Friedman, B., & Nissenbaum, H. (1996, July 1). Bias in computer systems. *Association for Computing Machinery*, 14(3), 330-347. <https://doi.org/10.1145/230538.230561>
17. Hardt, M., Price, E., & Srebro, N. (2016, December 5). Equality of opportunity in supervised learning. *arXiv (Cornell University)*, 29, 3323-3331.
<https://papers.nips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html>

18. Johndrow, J E., & Lum, K. (2019, March 1). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1). <https://doi.org/10.1214/18-aos1201>
19. Kamiran, F., & Calders, T. (2011, December 3). Data preprocessing techniques for classification without discrimination. *Springer Science+Business Media*, 33(1), 1-33. <https://doi.org/10.1007/s10115-011-0463-8>
20. Kobayashi, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, January 1). Fairness-Aware Classifier with Prejudice Remover Regularizer. *Lecture Notes in Computer Science*, 35-50. https://doi.org/10.1007/978-3-642-33486-3_3
21. Lee, N T. (2018, August 13). Detecting racial bias in algorithms and machine learning. <https://doi.org/10.1108/jices-06-2018-0056>
22. Mujtaba, D F., & Mahapatra, N R. (2019, November 1). Ethical Considerations in artificial intelligence -Based Recruitment. <https://doi.org/10.1109/istas48451.2019.8937920>
23. Obermeyer, Z., & Mullainathan, S. (2019, January 29). Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. <https://doi.org/10.1145/3287560.3287593>
24. Panch, T., Szolovits, P., & Atun, R. (2018, October 21). Artificial intelligence, machine learning and health systems. *Edinburgh University Global Health Society*, 8(2). <https://doi.org/10.7189/jogh.08.020303>
25. Pedreshi, D., Ruggieri, S., & Turini, F. (2008, August 24). Discrimination-aware data mining. <https://doi.org/10.1145/1401890.1401959>
26. Pfohl, S., Foryciarz, A., & Shah, N. (2021, January 1). An empirical characterization of fair machine learning for clinical risk prediction. <https://www.sciencedirect.com/science/article/pii/S1532046420302495>
27. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K Q. (2017, September 6). On Fairness and Calibration. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1709.02012>

28. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K Q. (2017, September 6). On Fairness and Calibration. *neural information processing systems*, 30, 5680-5689. <https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffb2d39ab038d1cd7-Abstract.html>
29. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020, January 27). Mitigating bias in algorithmic hiring. <https://doi.org/10.1145/3351095.3372828>
30. Rajkomar, A., Hardt, M., Howell, M D., Corrado, G S., & Chin, M H. (2018, December 4). Ensuring Fairness in Machine Learning to Advance Health Equity. *American College of Physicians*, 169(12), 866-866. <https://doi.org/10.7326/m18-1990>
31. Rajkomar, A., Hardt, M., Howell, M D., Corrado, G., & Chin, M H. (2018, December 4). Ensuring Fairness in Machine Learning to Advance Health Equity | *Annals of Internal Medicine*. <https://www.acpjournals.org/doi/10.7326/M18-1990?cookieSet=1>
32. Rouse, C E., & Kemple, J J. (2009, March 1). Introducing the Issue. *Princeton School of Public and International Affairs*, 19(1), 3-15. <https://doi.org/10.1353/foc.0.0022>
33. Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N A. (2019, January 1). The Risk of Racial Bias in Hate Speech Detection. <https://doi.org/10.18653/v1/p19-1163>
34. Speicher, T., Heidari, H., Grgić-Hlača, N., Gummadi, K P., Singla, A., Weller, A., & Zafar, M B. (2018, July 19). A Unified Approach to Quantifying Algorithmic Unfairness. <https://doi.org/10.1145/3219819.3220046>
35. Sweeney, L. (2013, March 1). Discrimination in Online Ad Delivery. *ACM Queue*, 11(3), 10-29. <https://doi.org/10.1145/2460276.2460278>
36. T, R K., Hemank, L., & Rayid, G. (2020, December 5). Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy. <https://doi.org/10.48550/arXiv.2012.02972>
37. Yuan., Ming., Kumar., Vikas., Ahmad., Aurangzeb, M., Teredesai., & Ankur. (2021, February 7). Assessing Fairness in Classification Parity of Machine Learning Models in Healthcare. <https://doi.org/10.48550/arXiv.2102.03717>

38. Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 3 (May-June 2018), 31–57. <https://doi.org/10.1145/3236386.3241340>
39. Zhang, B H., Lemoine, B., & Mitchell, M. (2018, December 27). Mitigating Unwanted Biases with Adversarial Learning. <https://doi.org/10.1145/3278721.3278779>
40. Zhang, K. (2022, February 24). Semantic Scholar. <https://www.semanticscholar.org/author/Kun-Zhang/2119017174>