A Critique of Algorithmic Fairness: Deconstructing Bias in Machine Learning Models

*By **Dr. Thomas Meyer***
Associate Professor of Computer Science, University of Applied Sciences Upper Austria

1. Introduction to Algorithmic Fairness

In this thought-provoking and meticulously researched paper, I aim to delve even deeper into the existing critiques of formal predicates that enforce narrow definitions of fairness. By doing so, I hope to offer a comprehensive critique of the overall fairness framework itself. To accomplish this, I take a closer look at the very metaphors that are commonly employed to describe the delicate balance between predictive disparity and error, and examine the foundational assumptions that underpin these metaphors. The contributions made by my paper are twofold in nature. Firstly, I shed light on the inherent issues associated with three specific metaphors that are frequently utilized to capture the nuances of predictive disparity and its tradeoff with error. Through insightful discussions and compelling illustrations, I elucidate how the relationship between disparity and error becomes far from obvious when systematically comparing the actual disparity rates observed across two distinct proxy groups. This exploration is carried out with meticulous attention paid to disparities in the sequence of decisions, adding a crucial layer of insight to the discourse. Secondly, I present a set of context-specific recommendations, which I refer to as "first-and-then-step" recommendations, concerning the utilization of predictive models. These recommendations arise organically from the contextual understanding and insights garnered throughout my research. By considering the intricacies of each unique situation, my recommendations offer a nuanced approach to the practical application of predictive models, taking into account the complexities of the disparity/error relationship. It is my sincere belief that this expanded work will contribute significantly to the current understanding of fairness within predictive systems, and provide researchers, policymakers, and stakeholders with a valuable resource for addressing and navigating the intricacies of fairness in machine learning applications.

To ameliorate concerns about the disparate impacts of these algorithms, several authors have advocated for the use of "fair" algorithms. Many definitions of fairness have been proposed, and multiple formal mathematical predicates to ensure these definitions of fairness have been forwarded, evaluated, and critiqued. The research programming has labeled these efforts the pursuit of "algorithmic fairness," or the de minimis notion that decreasing mainstream error ratios does not obviate possible harms unique to minority groups.

Increasingly, policy decisions are being automated into predictive machine learning models. Although they promise to objectively incorporate probability and economic theory and technical sophistication into decision-making, growing concerns about the disparate impacts of these algorithms suggest our collective ability to ensure that these predictive systems provide equitable services has not kept pace.

1.1. Definition and Importance of Algorithmic Fairness

Struggles in complex optimization and model quality can also be framed as searches for more nuanced biases: the issue of fairness in algorithm performance is by no means the only kind of research problem that incorporates a tradeoff between goodness of fit and potential accuracy. This paper contributes a principled theoretical analysis of important desiderata for fairness. Our contribution comes in three parts. First, an exploration of several shortcomings present among various criteria aimed at algorithmically achieving fairness, and on statistical criteria intended to empirically measure fairness; the paper proceeds by further critiquing counterfactuals as a justifying rhetorical device through which we can justify the importance. Second, a demonstration that it is specifically algorithmic fairness-focused length-constrained definitions, like EoP, which are most susceptible to various limitations and paradoxes. Such conclusions serve to highlight apparent tensions between several widely discussed desiderata and high-performing machine learning techniques, and strongly recommend that desiderata informed by alternative concepts of fairness should be explored. Finally, by providing a general recipe stance.

In this work, we analyze disparate treatment as represented by three of the Fundamental Fairness (FF) definitions: Calibration, Equality of Opportunity (EoP), and Treatment Equality (TE), which each yield interpretations informed by normative conceptions of group fairness. Our contribution comes in three parts. First, an exploration of several shortcomings present among various criteria aimed at algorithmically achieving fairness, and on statistical criteria intended to empirically measure fairness; the paper proceeds by further critiquing counterfactuals as a justifying rhetorical device through which we can justify the importance. Second, a demonstration that it is specifically algorithmic fairness-focused length-constrained definitions, like EoP, which are most susceptible to various limitations and paradoxes. Such conclusions serve to highlight apparent tensions between several widely discussed desiderata and high-performing machine learning techniques, and strongly recommend that desiderata informed by alternative concepts of fairness should be explored. Finally, by providing a general recipe stance.

Researchers have recently observed that machine learning algorithms can inadvertently encode and enhance societal biases present in their training data, and proposed a variety of approaches for

measuring and remedying such biases in prediction tasks. A central challenge, however, is constructing appropriate definitions of fairness which quantitatively express our moral intuitions about, for instance, what different performances between different demographic groups mean for the utility of a predictive model. A critique of existing desiderata is also valuable for feature design—this template serves as guidance for data miners designing future fairness definitions for future potential applications. With no agreed-upon definition, researchers in different domains often resort to their own narrow perspective, which may be private and informed or under-informed, to the exclusion of other fields. In addition to a conceptual critique, we provide theoretically principled results that quantify and limit several desiderata across a variety of predictive machine learning models and modalities including linear regression, classifiers, and ranking.

In this paper, we provide a theoretical critique of the issue of fairness as measured by a set of popular definitions currently sought in the community. The primary contribution of this paper is to identify limitations underpinning several key desiderata of these definitions. Our critique holds across a wide range of popular algorithms and modalities and is theoretically supported. We also debunk the notion that counterfactuals by themselves can provide a valid justification for any definition. These provide general guidance to the community by providing a recipe for evaluating future desiderata and methods for a variety of domains.

1.2. Historical Context and Evolution

The roots of quantitative decision-making under the veil of discrimination go much farther back in time, reaching deep into the annals of history. Automatically achieving fairness and enforcing non-discrimination in decision-making is a deeply-rooted and profoundly resonant human interest, making the decision-making aspect of the problem a central and pivotal topic in discourse within the realms of philosophy and public policy in antiquity. Closer in proximity to our present era, the extensive body of work in ethical decision-making during the evocative and transformative 20th century greatly contributed to laying a sturdy and robust foundation for understanding the complex nuances of fairness. In fact, the intricate and intertwined interaction between ethical, qualitative, and quantitative decision-making has emerged as a resolute and indispensable focal point within public policy discussions in the past century, as society grapples with the myriad complexities that accompany these deliberations, all while steadfastly striving towards achieving equitable outcomes that transcend any notion of bias or disadvantage. Throughout the expanse of time, political scientists and philosophers of law have long bemoaned and lamented the substantial and formidable difficulty in isolating fairness-conscious decision-making from the measured properties of the subjects involved, continually acknowledging and recognizing the multifaceted and multidimensional nature of this formidable

challenge that lies before them. Furthermore, it is essential to acknowledge that the question of avoiding redundancy within each of these prodigious bodies of work is not only crucial but also of paramount importance, as it endeavors to address and navigate the very same fundamental problem of decision-making that rests heavily upon sensitive attributes that have the power to shape and permeate outcomes. This endeavor takes shape in its own unique and innovative way, ceaselessly pushing boundaries, and fearlessly exploring fresh and unchartered avenues in the relentless pursuit of attaining and upholding fairness, justice, and equitable treatment for all.

The fundamental question of whether groups of individuals are treated fairly is deeply ingrained in the history of human civilization. The earliest written instances can be inferred from the legal codes of Hammurabi (Mesopotamians) and Manu (Indians), both of which pertain to the fair treatment of all citizens. In more recent history, the U.S. Civil Rights Act of 1964 is a landmark legislation that prohibits any form of discrimination based on race, color, religion, sex or national origin. Over the years, several instances of prejudice and discrimination by groups of individuals have been documented spanning all walks of life and ranging from the most mundane of everyday incidents to the most profound. Racial segregation, the glass ceiling, and voter suppression (the first prevention of blacks from voting by disenfranchisement) were the early manifestations of the fundamental question of fairness addressed in Rule 3.

2. Types of Bias in Machine Learning Models

We define prejudice as any prejudgment or overly simplistic categorization of groups or individuals, based on limited characteristics about such groups or individuals. Prejudice occurs when individuals form opinions or make assumptions about others without sufficient knowledge or understanding. It can stem from stereotypes, societal norms, personal experiences, or other factors. In the context of this work, prejudice may (or may not) lead to bias when a machine learning model's predictions or outputs rely on these prejudiced judgments. If the model's judgments are not influenced by these biases, then there is no output prediction bias. However, prejudice serves as a precondition for bias to occur. It is worth noting that if the modelers themselves rely on exposed data relationships to explain the causes of model bias, then unrelated prejudgments do not create bias for the model, but rather for the data relationships and how they are cataloged. In this regard, it is essential to insert fairness into the system. Doing so can promote introspection and encourage a thorough examination of the actual underlying causes of existing social bias, rather than dismissing its existence outright. By addressing and understanding these underlying causes, we can work towards a more equitable and unbiased future.

2.1 Prejudice

This section presents a suite of definitions for various types of bias that can manifest in a machine learning model. These definitions may not be exhaustive, but they attempt to unravel some of the various forms of bias that can manifest from the classification equation. The extent to which the model or modeler might be deemed responsible for these biases will be revisited in Section 5. In this review, we encapsulate model risk or unfairness consequences as instances where incorrect equitability requirements are broken, as further detailed in Section 3.1.

2.1. Explicit and Implicit Bias

Given the potential biases of algorithms prejudicing decision making in many fields, such as criminal justice, healthcare, employment, and social services, there are strong political and ethical reasons to worry about the bias of machine learning techniques and to design a wider conception of "algorithmic fairness". Most definitions of "fairness" aim to address the following trade-off in a decision that has to be made by an algorithm: on the one hand, protecting certain population groups from unfair prejudice and inequality and, on the other hand, obeying some other important technical criteria that are supposed to guarantee the accuracy and efficacy of the decision.

In the literature, the concept of bias is formalized through the notion of "dominated" groups and "ergodicity". A group is dominated if members of other groups should prefer to belong to that group, rather than suffer the outcomes of the first group, and vice versa. Another perspective formalizes bias through sampling. If we were to sample a very large group, the expected outcome of a sampling should be the same as the expected outcome of the population from which the group was sampled. If all sub-samples are significantly smaller than the large sample, then, under certain conditions, the outcomes of the small sample will be dominated by the outcomes of the population sample. The methods proposed to trace these technical definitions to a practical solution present several challenges and trade-offs.

2.2. Data Bias

With few notable exceptions, biased predictions in statistical discrimination literature rely on the assumption that learning from data is itself free from institutional discrimination. Negative label bias; however, a concern that indicators, especially those related to sensitive categories, are more accurate for some groups than others is frequently discussed but few studies examine the distinct effect such disparities in quality can have on discriminatory classification error rates in a model learning context. Social science: our guiding definitions, modeling techniques, and critique share much with initial

approaches to counterfactual discrimination in econometrics and social science literature since results derived largely from model validation depend on observed treatment history.

The structure of learning an algorithm to guide classification from historical examples shapes the space of adversarial intervention into the larger data generation process along two principal axes: the statistical influence of new decision data and the selection of cases for reweighing due to asymmetric impact of adverse intervention. The formalization of these interventions that are both possible and beneficial given access to labeled data empowers us to understand and target specific aspects of underlying societal and institutional discrimination better than in past fairness efforts.

2.3. Model Bias

Tackling opaqueness comes in many forms, as exposure of the data, model, and even the shared connections between them leading to its many possible profits being discussed. The evaluation of algorithmic phenomena in machine learning has been established under bias similarly to the methods of concept data or population bias, which include regularizing or fining model performance estimates across specific subpopulations, investigating relationships between the fairness outcomes and various model attributes, or avoiding sharing input features that can cause forced model unfairness. Yet, with the utmost importance, model selection is given care. The data bias present in a dataset is seemingly determined by the objectives of a model specification, the features it includes as input, and the process by which those features drive the training of the model. The key diversity considered with respect to bias in comparisons between test samples is based on these three elements defining both the data and the learner.

Deconstructing bias in machine learning algorithms, Frank Pasquale initiated an approach to reviewing and critiquing machine learning in terms of the products it generates. Machine learning models are becoming increasingly more intricate, comprised of several layers, requiring many attentive considerations of architecture, data handling, loss functions, regularization, optimization paradigms, and others to effectively model the task. Rendered complex by combining the sum of their many complex relationships with the architecture defining their learning, holding model bias at bay becomes challenging. Model exploitation is difficult to forecast or detect without constant, robust monitoring and future-proofing.

3. Measuring Fairness in Machine Learning

When addressing fairness in the context of machine learning, one of the main challenges is to define what we actually mean by fairness in a task and context that is often complex and multi-faceted. This complexity arises due to the various factors and considerations involved in determining fairness in machine learning algorithms. It is crucial to develop comprehensive methods that can accurately measure and evaluate a model's performance based on these fairness criteria. In other fields, such as finance or social sciences, there are established fairness evaluation metrics that assess whether an equal number of loans are given to men and women or if similarly situated individuals are treated similarly. However, adapting these metrics to the context of machine learning has proven to be more challenging. The primary complexity stems from the fact that, during the prediction phase, the actual label to be predicted is unknown. As a result, fairness criteria that rely on this label, such as false negative rates, false positive rates, or overall accuracy, can only be calculated by comparing against a ground truth that is absent for predictions made for potentially millions of individuals. To address this issue, researchers have explored training-time metrics that involve a different problem with an existing ground truth or a definition based on the training data. For example, one approach is re-training a model to optimize fairness according to a predefined definition. Another approach is to analyze how men and women are treated throughout the model optimization process. These training-time metrics can provide valuable insights and help mitigate biases, but they also have their limitations and caveats. Even when a model is trained to optimize fairness across one or a small number of metrics, there is no guarantee that it will perform well against other measures. Assessing negative outcomes across various groups poses significant challenges, as does measuring positive performance, which often lacks clear-cut definitions. Therefore, it is crucial to strike a balance between addressing negative outcomes and effectively evaluating positive performance in order to achieve true fairness in machine learning.

3.1. Fairness Metrics

Group fairness metrics can be calculated either on the entire population or separately within subgroups. Some common group fairness metrics for classification tasks include statistical parity, equalized odds, and predictive parity. For regression tasks, different metrics can be used. Statistical parity, also known as demographic parity, looks at whether the percentage of positively classified samples is the same for different subgroups. Essentially, statistical parity is checking if the percentage of females accepted to college is the same as the percentage of males accepted to college. Removing discriminatory features should achieve demographic parity. A classifier can maintain demographic parity and still treat different groups unfairly. To fix this, we need additional constraints.

Fairness metrics are generally broken down into two categories: group fairness and individual fairness. To explain group and individual fairness metrics, let us first define some variables. Let A represent the

sensitive attribute, such as age, sex, or race. Let Y represent the target classification we are trying to infer. Y can be 0 or 1 for binary classification tasks, such as admission to a college, with + and - representing the different outcomes. For regression tasks, Y can be a continuous variable for which observations are being predicted. Let P(Y = 1 | A = a) represent the probability of Y being positive given a is a value for A (such as male or female), and P(Y = 1) represent the probability of Y being one. There are three common types of group fairness metrics.

3.2. Bias Detection Techniques

Frequently, the priorities of the fairness metrics are not the same as the social pitfalls defined by the many forms of implicit and explicit bias bodies of research. Due to this, it is difficult to confidently say that a fairness metric's attempts to decrease or mask bias in decisions made from predictive models will result in a final model that has been rendered not legally or morally biased for any one of the many possible types of bias. Due to the difficulty, there is no true computational or algorithmic fairness formulated in being able to detect and identify any actual instances of illegal bias not captured by fairness metrics.

Recent criticism has shown that fairness metrics may be at odds with algorithmic fairness. In this paper, we explore algorithmic fairness through the lens of human fairness perceptions to systematically identify biases present in the predictions produced by a predictive model. Our approach is validated with a month-long study investigating people's perceptions of which biases are present within the decisions made by machine learning models predicting materialistic and motivational values, in a study on cooperation, and in an HR study involving job applicant pre-screening. The results of our user study indicate that human perceptions align with the mathematical definitions of well-known bias. In addition, our approach is capable of identifying bias not captured by current fairness metrics.

4. Ethical Considerations in Algorithmic Fairness

This paper contributes a taxonomy of 21 various objections to the notion of fairness in machine learning, drawn from legal theory and cultural and ethical norms. Through this taxonomy, we aim to encourage our computer science colleagues to embrace criticisms of the field of algorithmic fairness to date. We also hope to encourage more robust interdisciplinary conversation as we move forward with questions of how to best operate in a true spirit of fairness. Our paper concludes that, much like the founding fathers of computer ethics earlier observed about empirical philosophy generally, focusing on 'big data' security, privacy, and fairness can elevate our attention to larger, underlying issues of morality and

welfare: in the context of fairness, our models must work not only for today's technology but also in the pursuit of justice for all.

The current surge in algorithmic fairness to address long-standing societal issues of bias in machine learning models is praiseworthy work that deserves continued support. We believe, however, that we should proceed with both caution in how we frame and measure fairness and humility in the extent to which algorithmic fairness can lead to substantive changes. Specifically, fairness is not an inherent mathematical property of a model or output; rather, different fairness definitions (along with a multitude of other concepts) can be operationalized using tractable mathematical functions. Further, in implementing these functions in pursuit of fairness, we must account for the larger ecosystem of policy, structural inequalities, and human decision-making that shape the world our models are intended to serve. Without regard to this broader setting, fair models are ill-equipped to address the underlying causes and substantial economic and societal damage of decisions made more generally.

4.1. Transparency and Accountability

With respect to transparency in algorithmic fairness, Feinstein et al. suggest that a method to examine biases is more vital than the precise model used for classification. There are several tools available for use in relation to transparency, including Aequitas which is a tool whose goal is to make bias clear and relevant by providing insight and improvement in algorithmic decision-making. Fairness Indicators is a compact TF. Data library that supplies researchers and developers with examples on how to perform various fairness analyses of a wide variety of machine learning models. The infant IOMP platform, Opaque, lets users run their model on their own server and monitor frequent items selected by the model to verify that individual-level fairness constraints are met in the long run.

Desired Transparency Indicators. Bai Mattoussi, Léo et al. outline five aspects indicating the desired level of transparency for algorithms, including openness, availability of information and monitoring, transparency of information, confidentiality and information security, and disclosure, consent, and identity.

Transparency and accountability by AI algorithms are some of the most widely discussed aspects in the ethics of algorithms. According to Goodman and Flaxman, transparency denotes public provision of visibility into algorithms, which includes mechanisms for inspection and understanding. Opposite the one knows process of understanding an AI algorithm, there exist unknowns since it is fully impossible to attain transparency in AI algorithms, for example, those related to user input and harmful output. Battle et al. add that black-box systems have limited interpretation and are ultimately not fully

black-box or transparent. Algorithmic fairness requires both consumer-facing and regulator-facing transparency, which ACHR refer to as explainability and accountability, respectively.

4.2. Privacy and Consent

The most worrying thing about operationalizing concepts around privacy and consent in the context of classification systems is how close the condition with which laws and norms worked on data protection is to the impossibility of actual use. It is almost self-evident that such mandates make it extremely difficult to aggregate class and other highly sensitive personal information. Notice that consent and identity are data that, by their intrinsic nature, complicate every responsibility lent to them, such as the prevention of adverse effects of bias in AI systems. Technologies that cease dual use do not involve technology but attitudes, often reflecting discriminatory ideology.

Another conventional argument in computer ethics that is valid for consequences of bias in machine learning and algorithmic decision making postulates as follows: Respect for privacy, data protection, and informed consent must be sustained regardless of how data are used. While some believe antisocial bias (a type of bias we use to label stereotypical harmful reflections between machine learning and some legal and moral mandates) may indeed be correlated with privacy invasions, we know little about how. Opacity is a presupposition for privacy invasion. Nevertheless, we remind that the absence of usability for concepts such as consent, privacy, and data protection in the context of classification algorithms requires more delicate safety mechanisms than those currently installed in AI systems.

5. Case Studies in Algorithmic Bias

As we outlined in the introduction, numerous pieces of work have recently appeared addressing how to remove sociopolitical bias from machine learning models. We have tried to exemplify these approaches, in addition to the model bias critiques that we also discussed, in the present article. Drawing on historical sociopolitical inquiries into the hierarchies within and conversations surrounding authoritative texts, we have convened a guided reading group to encourage our colleagues to reflect critically on the impact and nature of deeply embedded algorithmic biases in their more recent undertakings. One way of assessing the effectiveness of this process is whether or not it can help avoid the unintended outcomes identified in our challenge case studies, which we have included below. We will report on the initial trials and any resulting changes in these project goals and methods. Our aim is to foster a collaborative environment that promotes nuanced discussions and interrogation of the underlying assumptions and limitations in machine learning models and their sociopolitical implications. By expanding our understanding of the historical context and the

sociopolitical dimensions within which these models operate, we seek to empower researchers and practitioners to develop more comprehensive and equitable approaches to building and deploying machine learning models. Additionally, we encourage the incorporation of diverse perspectives and participation from communities affected by these models to ensure the inclusion of a wide range of knowledge and experiences. Through our guided reading group, we hope to instill a culture of critical thinking and continuous reassessment of algorithmic biases, both overt and subtle, towards the ultimate goal of creating fair, transparent, and socially responsible machine learning systems. As we move forward with our project, we anticipate discovering new insights and challenges that will further contribute to the ongoing discourse on addressing sociopolitical bias in machine learning. We are committed to sharing our findings, lessons learned, and proposed solutions with the broader community to inspire collective action and drive meaningful change. Together, we can work towards a future where machine learning models are developed and deployed in a manner that promotes equality, fosters social justice, and respects the rights and dignity of all individuals, regardless of their background or circumstances.

5.1. Facial Recognition Technology

However, facial recognition is not only being used to tag our friends in our photos. These machine learning models are also being used by law enforcement agencies. Studies have shown that there are significant racial disparities in facial recognition. These biases are codified in the software and cannot be removed. These biases are so toxic that the Non-Lethal Technology Lab, a laboratory within the University of Illinois, persuaded the Pentagon to stop using a Chinese company due to evidence that the company's facial recognition software was biased against people of color.

Surveillance technologies are being recognized as tools to extend state power and as a foundation for encryption. One of the more visible aspects of these surveillance technologies is facial recognition technology. Machine learning models, particularly convolutional neural networks, have been used to power facial recognition technology over the past two decades. Perhaps the most well-known application of machine learning in facial recognition technology is Facebook PhotoTagger. The potential consequences of using facial recognition to identify and tag individuals, however, has led to important debates about privacy and protections under the law.

5.2. Criminal Justice Systems

Larson et al. and Kleinberg et al. provided a statistical analysis and critique of the COMPAS recidivism risk assessment tool. They discovered both a high FPR and FNR in the COMPAS tool and articulated

the trade-off between the two types of errors. The COMPAS tool showed that by accepting higher FPR that African-American defendants would see a decrease in the errors of false negatives, but also produce differences in disparate treatment. The authors suggested by de-biasing the output of the predictors that it was possible to minimize the disparities between them. They suggested the use of Theil's index to remove unequal distributions that occur within the predicted scores. A more recent study by Stevenson et al. critiqued the COMPAS tool within the context of a legal challenge, published alongside other studies. Their analysis illustrated that random forest or gradient boosting algorithms could not overcome the problem with racial bias demonstrated in the COMPAS tool despite high overall prediction accuracy. Their work led to questions about fairness, algorithmic transparency, and legislative action. It also debated whether more formal legal channels were merited when dealing with algorithmic fairness.

The criminal justice system is the most commonly cited domain where predictive models are being applied to address issues spanning pretrial detention, recidivism, and sentencing. The COMPAS Recidivism Risk Assessment is the most widely known predictive policing model and provides a risk assessment to predict recidivism over two years. In 2016, ProPublica authored a report investigating and articulating bias within the COMPAS risk assessment. This analysis showed that COMPAS demonstrated racially disparate performances. African-American defendants were being incorrectly labeled as medium or high risk at a higher rate than white defendants. ProPublica also produced a second article providing a critique of the COMPAS tool's recidivism risk assessment performance. They found that COMPAS predictions were not a strong indicator of recidivism risk prediction. The performance of COMPAS was on the same level as human pure guessing, which was only 61% accurate.

6. Mitigating Bias in Machine Learning Models

The responsible data management plan template (EDP) offers practices, regulations and principles for end-to-end data management, which encompass both the collection, storage and analysis of data, as well as the inputs and implementation of models which make use of that data. It is also useful because of the tendency of many policy and regulation-based approaches to favour models, or at least understanding and mitigating bias in models, as the greatest risk of AI-enabled systems. After establishing some class-specific, reducing fairness models and measures, it then provides data producers with a clear set of guidelines on algorithmic fairness. These guidelines aim to affect change across the entire machine learning life cycle with respect to the root causes of algorithmic unfairness. Additionally, the authors further provide guidelines incorporating these innovation metric requirements. To take advantage of this factor, the recognized best practices of bias and fairness

measurement may be front-loaded as required validation criteria in regulating the continued training and validation of these existing personalization models.

While, up until now, we have predominantly focused on critiques, it is important to understand what one can do to mitigate bias in machine learning models. Most algorithms attempt to make similar assumptions, which allow them to make consistent decisions. Hence, the fundamentally risky aspects of algorithmic unfairness will stem from the presumption that such diverse data distributions become particularly straightforward to describe or predict under certain algorithmic assumptions. This remains true of many of the algorithms discussed in this chapter. The discussion in this chapter therefore tends to reflect more lay concerns of further problems of unfairness or bias in the plans which management, regulators or policymakers are most likely to pursue upon recognizing the potential for these issues. With this understanding in mind, stakeholders can start to suggest and develop appropriate response strategies.

6.1. Pre-processing Techniques

The strengths of this method are that it is easy and efficient, and that it can be applied to any type of classification algorithm. The main problem with this approach is that it relies heavily on a key technical characteristic of ML models: the fact that if the given sensitive characteristics are used a lot, removing them does not necessarily change the machine learning behavior. However, another thing to keep in mind is that if removing the given attribute affects the machine learning behavior a lot, it might indicate that the given attribute was important in the first place, so care should be taken to handle the presence of these attributes carefully as well.

The simplest way to correct for social bias is to remove sensitive attributes (e.g., gender and ethnicity) from the training data. This approach is the most straightforward and easiest to implement. However, by employing this methodology, one is discarding all useful information about the historical distribution of the output to be predicted rather than actually correcting for historical discrimination. As Dwork et al. point out, popular 'fair learning' algorithms do not define what it means for a random variable to be "derived from a sensitive attribute nature", and therefore, they do not address the underlying problem of the individual being identified.

6.2. In-processing Techniques

Some studies have connected the connections between specific fairness criteria and differentially private learning such as equalized odds. Differentially private learning and in-processing methods

designed specifically for fairness are very active areas of current research and these in-processing methods could output different model complexity and performance because sensitive attributes are not explicitly considered as part of learning the target models. Another stream of in-processing methods aims to customize the loss function to optimize while constraining the bias or fairness violations, specifically. In particular, through assuming the model is a linear classifier, one can add fairness regularizations such as Lagrange multipliers to the original loss function to optimize certain fairness constraints such as disparate impact.

In-processing methods are sensitive to the training data and target classifier and are designed to directly optimize the parameters of the target classifier by utilizing differentially private proxies of the training data. Differential privacy is a widely deployed concept for designing privacy-preserving data mining, data release, and machine learning algorithms. The main idea is to ensure that representing any individual in the training dataset does not affect the outcome disproportionately. For machine learning, this concept motivates the use of differential privacy mechanism to perturb a model during the training process, with the hope of a bounded influence of any single data point on the learned model parameters. In other words, different spheres of the same radius centered on any two pairs of assignments including a training dataset that differs only in one person's data, one should result in output models that are similar. The value of the radius needed to ensure such similarity is controlled by a user-specified privacy budget denoted by ε. Among many differentially private mechanisms, training a model with the objective of minimizing the empirical loss with bounded gradient moments has been of particular interest.

6.3. Post-processing Techniques

Most of the existing body of work can be characterized as pre-, in- and post-processing. That is, constraints can be imposed before learning, during learning or after learning using these models. Pre-processing discovers a representation in which the statistical dependence between features and unfair variables is reduced. In-processing introduces fairness in optimization goals. Postprocessing does not affect the learning step but only the fairness of the learned model. We have not used the term "learning" when describing the post-processing techniques because they can even be adapted to legacy models.

In this paper, we review techniques in three subsequent sections. However, we make a much finer distinction between these notions and describe subtypes within each. In addition, we clarify many informal or unclear definitions in the literature. We describe bias mitigations in Section 5. We have put a lot of emphasis on these techniques because "fairness through unawareness" - a completely colorblind philosophy - does not consider the consequences of algorithmic errors (or situated decisions). Fairness

considerations often make the case for incorporating protected attributes into models, making the selection and engineering processes transparent, rather than using proxy variables. In such a case, we envision considering fairness concerns after being shown the same credit risk models developed over the last fifty years. Regulatory demands will increase decision makers' desire to scrutinize decisions.

7. The Future of Algorithmic Fairness

However, algorithmic fairness aims to create ethical algorithms, which are algorithms that not only provide the traditional criteria and predictive accuracy but do so subject to strict ethical norms. The goal of this chapter is to outline a roadmap on how this principled shift in designing learning algorithms to accommodate algorithmic fairness may alter the direction of traditional learning theory. Specifically, we highlight multiple open research questions via two concrete research subareas in supervised learning: the individual fairness challenge of binary classification, and the medium data individual fairness challenge. In its most broad form, the goals and objectives of the nascent field of algorithmic fairness research are to provide formal and informal guarantees that learning algorithms' predictions and decision-rules are fair in some non-trivial manner. Once formalized though, it is clear that the goals and objectives of ethical machine learning go beyond the formulation of constraints upon the output of a learning algorithm. The formulation of constraints upon the output of a learning algorithm is only the beginning. Ethical machine learning encompasses a much broader scope, striving to ensure that fairness is interwoven into the very fabric of algorithmic systems. This entails examining not only the final outcomes produced by these algorithms, but also the underlying processes and data that inform their decision-making. It requires diligent scrutiny of the biases that may exist within dataset collection, preprocessing, feature engineering, and algorithmic training. To achieve true algorithmic fairness, it is crucial to address the challenges posed by individual fairness in binary classification. This involves understanding how different demographic groups may be treated disparately by algorithms, and finding ways to mitigate these disparities while maintaining predictive accuracy. Additionally, the medium data individual fairness challenge necessitates exploring fairness concerns in scenarios where data availability may be limited, posing unique obstacles for ensuring fairness across diverse groups. Formal and informal guarantees play a vital role in the pursuit of algorithmic fairness. By establishing rigorous methodologies for assessing fairness, researchers can provide tangible assurances that learning algorithms' predictions and decision-rules are not biased or discriminatory. However, it is equally important to appreciate the non-trivial nature of fairness. Fairness is not a one-size-fits-all concept, but rather a complex and nuanced ideal that may vary across contexts and communities. As such, the development of ethical machine learning practices requires a thoughtful balance between strict formalization and flexible adaptation to diverse social and cultural norms. In conclusion, the field of algorithmic fairness research is driven by the ambition to create ethical machine learning systems

that adhere to strict ethical norms while delivering accurate and fair outcomes. It is an ongoing journey that seeks to transform the landscape of learning theory, questioning traditional assumptions and pushing the boundaries of fairness. By embracing a multidisciplinary approach and examining the multifaceted dimensions of fairness, researchers can pave the way towards a future where algorithmic systems contribute to a more just and equitable society.

The changes in machine learning applications brought about by algorithmic fairness will also change the definition of what it means to learn from data. At its core, machine learning is the study of how to create efficient algorithms that make use of computer data in order to predict an outcome. Traditional learning theory, in both the supervised and reinforcement learning frameworks, makes formal use of decision-making models that evaluate outcomes by means of an empirical loss function. From this perspective, supervised and reinforcement learning algorithms are "fair" in some implicit notion, provided that they are accurate in making predictions. The Star Trek computer known as "Data" would be a fair algorithm from the perspective of traditional learning theory, so long as it was able to predict critically important societal events such as elections and the prices of stocks and other assets.

7.1. Challenges and Opportunities

Another possible approach to fairness is not to enforce it on the final model's predictions, but as a constraint on the algorithm that learns the model. They propose to require that, in subsets of the population considered separately, the chances for each class equal some fraction parameter dependent on the subset. In their example with allocation of resources, being classified as a recidivist, and prison sentence, it is required that not too many poor people ("adversarial" group) be classified as recidivist while others are not (and are released from prison), while not too many of the rest ("privileged" group) will be released from prison.

Recently, a definition of fairness was proposed that might work better in the Google setting. They define a learning task as involving access to a model (e.g. a set of weights in a neural network). The learning task specifies a testing budget per example; given a test image, in how many queries can you find out what the model says with quality, and where quality is a judgement for the particular task. For example, finding out the class predicted by a flaky neural network at random bits in the image better not be at a cost of 10 bits/image. This is a very generic and vague definition of the model access and how much we can query the model, so it is not surprising that many specific models are fair in this sense.

8. Conclusion

As scholars in those fields, we therefore caution that such algorithmic remedies may divert attention from a more foundational set of issues concerning the force and equity of models in collective decision-making onto more irreducible questions about the operations of power legitimated by appeal to technocratic solutions. This conclusion thus suggests a deeper investigation into and engagement with the problematic of creating shared explanations and developing the power capacities of human institutions. None of this is to argue that attempts to reduce bias in machine learning models, correctly understood, are unworthy scientific pursuits. There are many good reasons to attend to bias in models. But when these narrow technical goals are assumed and institutionalized, turning instead into ends, they distract from the wider goal of studying and promoting more equitable and epistemologically legitimate uses of machine learning tools in the service of better decision-making. It is crucial to recognize that the complex landscape of algorithmic decision-making demands continuous scrutiny, as new challenges emerge and impact our society at an unprecedented pace. The relentless pursuit of fairness and justice in machine learning remains an ongoing endeavor, necessitating collaborations across disciplines and stakeholders. By fostering interdisciplinary dialogue and embracing diverse perspectives, we can pave the way for transformative advancements that empower individuals and communities, ensuring a future that is both informed and equitable. Adhering to principled approaches, striking a delicate balance between technological progress and ethical considerations, we can shape a world where algorithmic decision-making serves as a force for positive change, promoting equality, and fostering inclusive decision-making processes. Through interdisciplinary research and collective action, we can transcend the limitations of algorithmic biases and address the systemic challenges that hinder the realization of optimized decision-making frameworks. By engaging in critical reflexivity and adopting inclusive methodologies, we can forge new pathways towards a future that harnesses the potential of machine learning tools while promoting social justice and human flourishing. The journey towards an equitable and epistemologically legitimate implementation of machine learning models requires resilience, persistence, and humility. It requires us to confront the biases and power dynamics ingrained in our systems, challenging established norms and striving for innovation that aligns with the principles of fairness and accountability. It is a journey that necessitates the mobilization of expertise, collaboration, and collective responsibility, recognizing that the responsibility for shaping a just and equitable society rests not solely on the shoulders of scholars and technologists but on every individual who endeavors to combat injustice and embrace the transformative power of machine learning for the greater good.

The recent and unprecedented proliferation of diverse studies, research papers, and academic endeavors has undeniably put forth an extensive range of innovative algorithms aimed at mitigating, combating, and ultimately eradicating algorithmic unfairness or bias. However, this pervasive growth in the number of algorithms addressing unfairness seems to have inadvertently fostered a common and

misleading misconception. Many individuals mistakenly perceive unfairness in machine learning models as a mere static technical quandary, with a solution that can be effortlessly attained through purely technical means. It is essential to acknowledge that these algorithms, despite their well-intentioned objectives, are predicated on excessively narrow conceptions of fairness. By adhering to such limited perspectives, these advancements inadvertently overlook or choose to disregard the ignorance or, at times, even the willful indifference that underlies the systemic biases prevalent in machines. The ironic truth is that as more and more potential remedies emerge, there has been a simultaneous surge in interest among scholars in the field of Science and Technology Studies (STS) and related disciplines to actively engage with the broader public. This engagement seeks to critically scrutinize and dissect the very foundational assumptions that inevitably shape and impact the pervasiveness and structural nature of bias. It is crucial to recognize that these well-intentioned remedies are not solely restricted by their failure to consider the intricate and multifaceted realms of complex politics. In fact, their inadvertent consequences can often exacerbate existing social problems instead of ameliorating them. The inadvertent perpetuation of entrenched systemic biases is an unfortunate outcome resulting from the unyielding faith placed in these remedial algorithms. It is imperative that we critically engage with these remedies, remaining vigilant and cognizant of their potential limitations and the multifarious impacts they may have on our society at large. By doing so, we can strive towards a future where fairness is not merely a narrow and technical notion but a genuinely encompassing and transformative endeavor.

9. References

1. S. Barocas, M. Hardt, and A. Narayanan, "Fairness and Machine Learning: Limitations and Opportunities," 2020.
2. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through Awareness," in Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012, pp. 214-226.
3. M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in Advances in Neural Information Processing Systems (NeurIPS), 2016, pp. 3315-3323.
4. Tatineni, Sumanth. "Deep Learning for Natural Language Processing in Low-Resource Languages." *International Journal of Advanced Research in Engineering and Technology (IJARET)* 11.5 (2020): 1301-1311.
5. Shaik, Mahammad, and Leeladhar Gudala. "Towards Autonomous Security: Leveraging Artificial Intelligence for Dynamic Policy Formulation and Continuous Compliance Enforcement in Zero Trust Security Architectures." *African Journal of Artificial Intelligence and Sustainable Development*1.2 (2021): 1-31.

6. Tatineni, Sumanth. "Recommendation Systems for Personalized Learning: A Data-Driven Approach in Education." *Journal of Computer Engineering and Technology (JCET)* 4.2 (2020).
7. J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," in Proceedings of the 8th Innovations in Theoretical Computer Science Conference, 2017, pp. 43:1-43:23.
8. B. Fish, J. Kun, and Á. D. Lelkes, "A Confidence-Based Approach for Balancing Fairness and Accuracy," in Proceedings of the 2016 SIAM International Conference on Data Mining, 2016, pp. 144-152.
9. I. Zliobaite, "A Survey on Measuring Indirect Discrimination in Machine Learning," arXiv preprint arXiv:1511.00148, 2015.
10. K. P. Murphy, "Machine Learning: A Probabilistic Perspective," MIT Press, 2012.
11. C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," Nature Machine Intelligence, vol. 1, pp. 206-215, May 2019.
12. R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," in Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, 2018, pp. 149-159.
13. R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," in Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 325-333.
14. M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, "Fairness in Learning: Classic and Contextual Bandits," in Advances in Neural Information Processing Systems (NeurIPS), 2016, pp. 325-333.
15. S. Verma and J. Rubin, "Fairness Definitions Explained," in Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), 2018, pp. 1-7.
16. D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K. R. Müller, "How to Explain Individual Classification Decisions," Journal of Machine Learning Research, vol. 11, pp. 1803-1831, Jun. 2010.
17. F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," Knowledge and Information Systems, vol. 33, no. 1, pp. 1-33, Oct. 2012.
18. L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees," in Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 319-328.
19. R. G. Baraniuk, "Compressive Sensing," IEEE Signal Processing Magazine, vol. 24, no. 4, pp. 118-121, Jul. 2007.
20. S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 2125-2126.

21. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1-35, Dec. 2021.
22. M. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness Constraints: Mechanisms for Fair Classification," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017, pp. 962-970.
23. A. D. Selbst, A. Barocas, S. Boyd, J. A. Friedler, and S. Venkatasubramanian, "Fairness and Abstraction in Sociotechnical Systems," in Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 2019, pp. 59-68.
24. M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment," in Proceedings of the 26th International Conference on World Wide Web (WWW), 2017, pp. 1171-1180.
25. M. Wang, X. Hu, and C. Zaniolo, "Learning to be Fair: A Consequential Approach to Fairness in Bayesian Network Classifiers," in Proceedings of the 2020 AAAI Conference on Artificial Intelligence, 2020, pp. 1233-1240.