Fair Machine Learning and Anti-subordination

Vijay Keswani

May 14, 2021

1 Introduction

The availability of large datasets and massive computing power has led to a surge in use of automated decision-making frameworks in a variety of domains; the list of applications includes (and is not limited to) healthcare, risk assessment, advertising, lending, and content moderation. However, the presence of biased datasets and/or use of misrepresentative models has had an impact on these automated frameworks as well; multiple studies have pointed out that prediction algorithms often propagate biases and negative stereotypes against historically-oppressed groups. Given the importance of the applications where automation is being employed currently and the possible lack of human oversight in their deployment, it is important that the learned algorithms not only stop the propagation of such biases but also make efforts in curbing the factors that enable such discrimination in first place.

To that end, the ideal goal of any algorithm that aims to be "fair" with respect to the demographics of the underlying population should satisfy the *anti-subordination* principle. The anti-subordination principle argues that, in the light of existing and historical biases against disadvantaged groups, decision-making policies should ensure that all groups are provided equal treatment and equal opportunities [Fis76]. Given that certain groups have already been subjected to marginalization in areas such as employment, education, housing, etc., simply ensuring current equal treatment of all individuals is not sufficient as it does not address the impact of historical biases. Infact, to satisfy the anti-subordination principle, it is often necessary to treat individuals with different protected attributes (e.g., race, gender) differently using policies such as affirmative action [HX20].

Fair machine learning, on surface, focuses on a similar idea of using protected attribute information of individuals to ensure that the learned algorithms provide equal opportunities to individuals from different groups [BHN17]. However, anti-subordination is often not the primary stated motivation behind these algorithms. In this project, I explore two settings where anti-subordination principle is not satisfied despite the use of fair classification. Section 2 provides a brief introduction to the field of fair machine learning while Section 3 provides an introduction to the anti-subordination principle. Section 4 and 5 expand upon two settings where fair classifier might not satisfy the anti-subordination principle; the first concerns the choice of fairness metrics while the second is on the choice of granularity and intersection of protected attributes.

2 Fair Machine Learning

The broad goal of fairness in machine learning (ML) is to investigate and mitigate the social biases of automated algorithms. Recent work in this field has shown that, due to either biased training data or shallow algorithmic design, automated predictions from ML algorithms used in real-world applications are often biased with respect to protected attributes (such as gender, race, etc.). The settings where such biases have been discovered include many "high-stakes" applications, such as predictive policing [Was18], healthcare [MDL$^+$13], and credit scoring [KMZ19]. Given the increasing prevalence of automation in our society, the study of disparate treatment by automated systems is timely and crucial.

To quantify the "unfairness" of an ML algorithm with respect to a protected attribute, one has to first define a metric that captures this unfairness. Commonly used metrics, such as statistical parity or predictive parity [BHN17], are usually inspired by legal precedents on anti-discrimination and affirmative action, such as Title VII of the Civil Rights Act of 1964 [Blu78] or the 80% disparate impact rule [Bid06]. However, there are often context-dependent incompatibilities between the legal and technical definitions of fairness metrics that make their usage questionable [XR19].

Formally, to define a fairness metric and its usage in ML, we need the following setup. Suppose we have a population of individuals whose features lie in a domain Ω. Let D denote the underlying distribution of the individuals in the population. Any feature vector $x \in \Omega$ represents the demographic and application-related information of an individual and can be used to predict a class label y corresponding to the individual. For the sake of simplicity, assume that the class label is binary, i.e., $y \in \{0, 1\}$. For example, in the risk assessment setting, x usually contains the demographic information of any given defendant and his/her prior criminal record; the class label in this case is whether the defendant recidivates or not. Along with a feature vector and class label, each individual also has a protected attribute z (from domain $\mathcal{Z}$) associated with them; this can correspond to gender, race, age, etc.

The goal of classification in machine learning is to construct a model that uses the features x to predict the class label y. Given a hypothesis class of classifiers $\mathcal{F}$, the standard practice to choose a classifier from the class that minimizes an expected prediction-loss function over the population (e.g. minimize expected prediction error); the chosen classifier is

$$\arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim D} L(f(x), y),$$

where $L : \{0, 1\}^2 \to \mathbb{R}_{\geq 0}$ is a point-wise loss function.

A classifier $f : \Omega \to \{0, 1\}$ is said to "unfair" with respect to protected attribute Z if the predictions from f treat the different groups defined by Z differently. With respect to protected attribute Z, suppose that the fairness metric $h : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}_+$ quantifies the disparate impact of classifier f; the larger the value of $h(f, Z)$, the higher is the disparate impact of f. A classifier f can be considered to be fair in this case if $h(f) \leq \varepsilon$, for some small $\varepsilon > 0$. To design fair classifier [BHN17], prior work usually aims to solve the constrained optimization problem of

$$\begin{gathered}\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim D} L(f(x), y)\\ \text{such that } h(f) \leq \varepsilon.\end{gathered}$$

2.1 Fairness Metrics

Research in fair classification has focused on many different fairness metrics. The choice of fairness metric is highly important towards the eventual goal of ensuring fair prediction but at the same time its appropriateness is often dependent on the application. A few examples of fairness metrics and their applications are listed below.

2.1.1 Statistical Rate

Statistical rate implies that the selection rates of all groups defined by protected attribute Z should be equal. In this case, $h(f) := |\mathbb{P}[f = 1 \mid Z = 0] - \mathbb{P}[f = 1 \mid Z = 1]|$ captures the extent of unfairness of f with respect to statistical rate.

This metric is useful in many lending settings. For example, suppose we want that the rate of success of a loan application should be approximately the same for men and women (say $z = 1$ corresponds to women); in this case, statistical rate would be an appropriate metric to capture the disparate impact of any loan application policy [LDR+18, BHN17].

2.1.2 Equalized Odds

Equalized odds fairness metric conditions on the true class label of the individuals while determining error rate disparity across groups. In particular, equalized odds is satisfied when true positive and false positive rates of all groups defined by Z are equal. Let $h_1(f) := |\mathbb{P}[f = 1 \mid Y = 1, Z = 0] - \mathbb{P}[f = 1 \mid Y = 1, Z = 1]|$ capture the difference in true positive rates and $h_2(f) := |\mathbb{P}[f = 1 \mid Y = 0, Z = 0] - \mathbb{P}[f = 1 \mid Y = 0, Z = 1]|$ capture the difference in false positive rates; then $h := h_1 + h_2$ quantifies the deviation from achieving equalized odds.

Equalized odds fairness metric is useful when class-specific error rate is important to measure. For example, in case of risk assessment, disparity in false positive rates would imply that one group is being subjected to higher unwarranted risk scores than the other, which can lead to significantly harsher treatment or penalties for individuals that are, in reality, low risk [Cho17].

3 Anti-subordination

The eventual goal of any policy that aims to reduce or remove discrimination from decision-making is to deconstruct the social hierarchies that enable such disparities. Proponents of affirmative action policies in fields such as employment or education have argued that decisions such as college admissions or recruitment should ideally be based on future potential of the individual rather than existing/prior data or scores that can potentially contain problematic historical biases. Similarly, in machine learning, algorithm designers have a responsibility to assess the future impact of automated predictions with respect to the protected attribute and not solely rely on improving accuracy over a given training dataset.

Anti-subordination indeed captures this goal quite succinctly. Anti-subordination, in the context of civil rights, "contends that guarantees of equal citizenship cannot be realized under conditions of pervasive social stratification and argue that law should reform institutions and practices that enforce the secondary social status of historically oppressed groups" [BS03]. A selection policy satisfies the anti-subordination principle if it does not have an effect of subordinating any group.

In the context of automated decision-making, anti-subordination principle can come into play in two ways: auditing for biases and mitigating them.

- Bias audit: Auditing a prediction algorithm for biases can be done in multiple ways. If the goal is to simply not use the protected attribute for prediction, then just analyzing the input features and their correlation with protected attribute is sufficient. However, this often does not provide a complete picture of possible biases that an algorithm can induce. For example, as we see in a later section, despite ensuring "immediate fairness" in selection policies, the decisions made using such policies can have the impact of further propagating the biases that they are meant to address.

  An anti-subordination approach to bias audit would involve looking at the future impact of current selection policies and whether they reduce the gaps between population groups that enable discrimination of one group relative to the other.

- Bias mitigation: Similar to bias audit techniques, anti-subordination principle requires that current selection policies that aim to address bias in their predictions should also ensure that the predictions have a positive impact on the historically-oppressed groups. An example of such a bias mitigation approach in real-world settings are affirmative action policies in employment and education [HX20].

In contrast to anti-subordination, anti-classification simply argues that the protected attribute information of the individuals should not be used for decision-making. In the context of fair automated decision-making, anti-classification is an easier principle to satisfy. This could either be due to feasibility of implementation of anti-classification principles using current methodologies or due to lack of data about future impact of current predictions [XR19] (although recent work seems to be address the second point [PZMDH20]). Nevertheless, anti-subordination is arguably a more important goal to accomplish in bias audit and mitigation [BS03], as highlighted in the previous paragraph.

Recent work on future impact of fair classifiers has analyzed the different kinds of impacts that fairness constrained optimization can induce. In certain cases the choice of fairness metrics or protected attribute granularity can lead to fair classifiers not having a positive impact on reduce possible future disparate impact. We look at two such settings where, despite the use of fair classification, the principle of anti-subordination is not satisfied. The two settings considered primarily highlight the importance of choice of fairness metric and the choice of granularity of protected attribute.

4 Importance of Fairness Metrics to Anti-subordination Principle

To highlight the importance of fairness metric in achieving anti-subordination, we present the example from Liu et al. [LDR+18] and argue how their analysis shows that fair classifiers can also cause subordination

Setup. Suppose that all individuals in the population Ω have a score in $\{1, \ldots, C\}$ associated with them. Say the population can be divided into two groups A and B, with A comprising g_A fraction of the population. Let π_A, π_B denote the distribution of scores of group A, B respectively.

One example of such score is the credit score with the groups denoting different racial or gender categories. Due to possible historical biases or measurement errors in data collection, the distributions π_A, π_B can be different from each other, with scores for one group being larger on average than the other. Such biases in credit scores have indeed been observed in real-world scenarios

The purpose of having these scores associated with the population is to use them to make future decisions. Suppose that the institution selects a policy $\tau := (\tau_A, \tau_B)$ to assign 0-1 outcomes to the individuals; $\tau_j : \{1, \ldots, C\} \to [0, 1]$ is the policy for group j, i.e., $\tau_j(x)$ is the probability the institution assigns positive outcome to an individual from group j with score x.

The institution can decide a policy τ based on different factors:

- Given a utility function $u : \{1, \ldots, C\} \to \mathbb{R}$, policy τ can simply be the one that maximizes the expected utility $\mathbb{E}[u]$ over the population;
- In case of desired fairness across group membership, policy τ can be the one that maximizes expected utility subject to fairness constraints. For example, if the fairness metric is statistical rate, then the constraint will be to equal selection rates for group A and B. Say $\beta_j := \sum_x \pi(x)\tau(x)$, then equal selection rates imply $\beta_A = \beta_B$.

Importantly, the goal of the analysis is to look at the impact of different policies on the group score distributions. For example, once again suppose that the scores represent credit scores and the institution uses these scores to decide loan applications. Depending on whether an individual whose loan application is successful pays back the loan or not, the group score distributions will change. To quantify the impact of decisions, we need to measure the expected relative change for each individual. Let $\delta(x) : \{1, \ldots, C\} \to \mathbb{R}$ denote the expected change in score of an individual with score x. Then the average impact on group j due to policy τ is

$$\Delta_j(\tau) := \sum_x \pi(x)\tau(x)\delta(x).$$

If $\Delta_j(\tau) > 0$, then the policy τ leads to improvement in scores of group j, while $\Delta_j(\tau) < 0$ implies that policy τ leads to active harm.

Results. Having described the setup of the analysis in [LDR+18], their results on statistical rate are briefly described below.

The main result of [LDR+18] is the following: *under mild assumptions on institution utility, there exists population proportions g_A for which policy τ which satisfies statistical parity causes active harm on group A.*

In other words, by being over-eager in assigning positive outcome to individuals from one group, the policy τ that tries to satisfy equal selection rate (statistical parity) leads to decrease in average group scores of that group. For example, in case of credit scores and loan applications, an equal selection rate policy τ can end up accepting loan applications of minority individuals who are unable to pay it back, leading to decrease in their credit scores and decrease in average credit scores of the entire group.

Discussion in the context of anti-subordination. As mentioned earlier, anti-subordination principle crucially requires that the selection policy not have the effect of sub-ordinating any group. In the context discussed above, despite the use of a classification policy that is fair with respect to statistical rate, the score distribution of one group is still significantly inferior and different than the other group. Considering the fact that the source of the difference in distributions is mostly likely historical biases and the lack of opportunities provided to one group compared to the other, the use of fair classification has no positive impact on addressing this bias. Infact, the results of [LDR+18] imply that in certain cases policy that ensures equal selection rates might exacerbate the bias and increase the difference between the group score distributions by being over-eager in assigning positive outcome.

This paper highlights that the choice of statistical rate as the fairness metric is not suitable in every setting if the goal is to achieve anti-subordination. The problem persists even with other fairness metrics; similar active harm regimes are also possible when equalized odds is used as the fairness metric for fair classification [LDR+18].

Papers addressing choice of fairness metrics. While anti-subordination is often not the stated motivation of fair classification, a number of recent papers have indeed take the future impact of selection policies into considering while designing them. Zhang et al. [ZKL20] analyze feedback models that take into consideration the impact of fair classifiers on the underlying population. Corbett-Davies and Goel [CDG18] also highlight the challenges that come along with various definitions of fairness. In the context of impact of current decisions on underlying distributions, the performative prediction approach also aims to construct classifiers that are stable and optimal with respect to the distributions that automated predictions can induce [PZMDH20, MPZ21].

The ideas explored in the above papers align with anti-subordination principles and can be used to provide a more comprehensive way of choosing fairness metrics and constraints.

5 Importance of Group Intersectionality and Granularity to Anti-subordination Principle

A policy is said to have disparate impact if there is inherent and statistically-significant bias in the treatment of individuals based on their gender, race, age, sexuality, national origin, etc. While a number of fair classification approaches can be employed to ensure fairness with respect to any of the above attributes, the impact of such approaches on sub-groups defined by intersection of these attributes can often be overlooked.

In the context of automated decision-making, consider the example of a classifier that aims to ensure equal selection rates for men and women and equal selection rates for Caucasian and non-Caucasian individuals. Even if the selection rate is equal across gender and equal across the binary race, it might not be equal across their intersection, i.e., equal for, say, Caucasian men vs non-Caucasian women. The following simple example from Kearns et al. [KNRW18] illustrates this point: suppose a classifier returns a positive outcome for 50% of Caucasian men and 50% of non-Caucasian women. In this case, selection rate across gender is equal and selection rate across race is equal; but clearly, this classifier has significant disparate impact with respect to the intersection of gender and race. Kearns et al. [KNRW18] highlight other non-trivial settings as well where supposedly fair classifiers can have disparate impact with respect to sub-groups that are not considered while designing fairness constraints.

Similarly, Buolamwini, Gebru [BG18] analyzed the performance of commercially available gender classification tools for images. The classifiers analyzed were pre-trained ones provided by Microsoft, IBM, and Face++; each classifier had an overall gender classification accuracy in the range of 85-95%. However, a gender-specific analysis showed that the accuracy was always relatively higher for images of men than images of women, with the difference in accuracy ranging from 8 to 20%. More surprisingly and importantly, an analysis with respect to intersection of gender and skin-tone revealed a much more disturbing picture. The accuracy for images of dark-skinned women was significantly smaller than other groups for all classifiers; the difference between accuracy for dark-skinned women and light-skinned men was as high as 34% in case of IBM classifier. While these classifiers are not supposedly trained to satisfy fairness constraints (although the details of their construction are proprietary), this paper further highlights the importance of studying fairness with respect to all relevant sub-groups.

Discussion in the context of anti-subordination. In the context of anti-subordination, once again the above examples and papers show that selection policies can only satisfy this principle if they strictly do not have the effect of sub-ordinating any group. Real-world applications of classification, such as in risk assessment or lending settings, are often expected to be able to provide accurate and fair predictions for individuals with different backgrounds and demographics. Nevertheless, not taking intersectionality into account can lead to further marginalization of intersectional subgroups that already suffer from historical oppression.

The problem extends beyond intersectionality as well. Consider a binary race attribute, e.g. Caucasian vs non-Caucasian or African-American vs non-African-American, is bound to leave disparate impact with respect to other subgroups, such as women of color, either unchanged or exacerbated. Even with respect to gender, binary categorization into male vs female can be problematic. A study conducted by Scheuerman, Paul and Brubaker [SPB19] showed that existing commercial facial analysis tools do not perform well for transgender individuals and are unable to infer non-binary gender, primarily because of focus of training on recognizing gender-stereotypical facial features.

Ensuring fairness with respect to only a high level division of these attributes is often necessary to ensure feasibility of the fairness constrained optimization problems. Furthermore, given a biased and limited dataset for training, additional constraints can make the task ensuring low error rate (compared to an unconstrained classifier) significantly more difficult. For example, Kearns et al. demonstrate that the theoretical infeasibility of auditing and mitigating fairness for some natural classes of sub-groups defined by popularly considered protected attributes. In the context of such possible intractability of fair classification, an obvious question arises as to whether there are settings where fair constrained classification is indeed preferable over unconstrained classification and a number of papers have explored this question.

Finally, to summarize, the goal of anti-subordination requires addressing the biases that enable discriminatory selection policies. If this goal is to be truly realized, the study of intersectionality and granularity of protected attributes in fair classification is absolutely necessary and crucial.

Papers addressing intersectionality. Beyond highlighting the difficult of ensuring fairness with respect to all relevant subgroups, Kearns et al. [KNRW18] also provide certain heuristic techniques to construct classifiers that satisfy such complex fairness criteria. Other papers have extended their results and explored intersectional and sub-group fairness in many different settings [BL19, FIKP20]. Once again, while it is not the explicitly stated motivation of these papers, the problems and solutions they explore do align with the goals of anti-subordination.

6 Conclusion

The examples discussed above demonstrate that fair classification does not always lead to anti-subordination. An inappropriate choice of fairness constraint or ignoring intersectional sub-groups can lead to additional harm towards marginalized groups.

Using automated frameworks to replace humans in decision-making pipelines is expected to be independent of extraneous factors that often influence human decision-making. However, the influence of such factors on datasets employed for learning has led to such biases creeping into machine decisions as well, painting a pessimistic picture that machine decisions might not infact be better than human decisions. An adherence to anti-subordination principle in fair machine learning can hopefully address such issues.

This project highlights just two issues that can affect whether a fair classifier satisfies anti-subordination principle. There are many other context-dependent factors that can be important towards this principle too such as presence of humans in the loop and inclusion of updated/additional features in future classification.

References

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[BHN17] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

[Bid06] Dan Biddle. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

[BL19] Avrim Blum and Thodoris Lykouris. Advancing subgroup fairness via sleeping experts. *arXiv preprint arXiv:1909.08375*, 2019.

[Blu78] Ruth G Blumrosen. Wage discrimination, job segregation, and the title vii of the civil rights act of 1964. *U. Mich. JL Reform*, 12:397, 1978.

[BS03] Jack M Balkin and Reva B Siegel. The american civil rights tradition: Anticlassification or antisubordination. *Issues in Legal Scholarship*, 2(1), 2003.

[CDG18] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[FIKP20] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE, 2020.

[Fis76] Owen M Fiss. Groups and the equal protection clause. *Philosophy & Public Affairs*, pages 107–177, 1976.

[HX20] Daniel E Ho and Alice Xiang. Affirmative algorithms: The legal grounds for fairness as awareness. *arXiv preprint arXiv:2012.14285*, 2020.

[KMZ19] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285*, 2019.

[KNRW18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

[LDR$^+$18] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[MDL$^+$13] Tracy MacIntosh, Mayur M Desai, Tene T Lewis, Beth A Jones, and Marcella Nunez-Smith. Socially-assigned race, healthcare discrimination and preventive healthcare services. *PloS one*, 8(5):e64522, 2013.

[MPZ21] John Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. *arXiv preprint arXiv:2102.08570*, 2021.

[PZMDH20] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[SPB19] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019.

[Was18] Anne L Washington. How to argue with an algorithm: Lessons from the compas-propublica debate. *Colo. Tech. LJ*, 17:131, 2018.

[XR19] Alice Xiang and Inioluwa Deborah Raji. On the legal compatibility of fairness definitions. *arXiv preprint arXiv:1912.00761*, 2019.

[ZKL20] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Long-term impacts of fair machine learning. *Ergonomics in Design*, 28(3):7–11, 2020.