Algorithmic Fairness: A Necessity Result

Zeynep Burce Gumuslu

AI Alignment Seminar, December 2024

1 Introduction

(?, ?) (?, ?) (?, ?) (?, ?)

Predictive algorithms have been playing an increasingly prominent role in decision-making processes that significantly impact the well-being and prospects of individuals and groups. These algorithms are extensively used in fields such as criminal justice, healthcare, banking, employment, and education, where their outputs shape critical decisions ranging from sentencing recommendations to hiring outcomes. While these systems promise efficiency and consistency, concerns about their potential biases and unfairness have garnered substantial attention. These concerns are not merely hypothetical but are substantiated by numerous case studies revealing that real-world algorithmic systems often exhibit biases against somef socially salient groups.

Arising from such concerns, algorithmic fairness is a field of study that seeks to evaluate decision-making procedures based on their fairness, particularly in relation to what are often called *sensitive traits*, such as race, gender, ethnicity, and religion. A central challenge in this field is identifying appropriate criteria for assessing fairness in rule-based procedures. One influential approach involves defining statistical criteria for algorithmic fairness. These criteria are considered to be necessary conditions that the outcomes of an algorithm's predictions must meet for the algorithm to be deemed fair. Prominent examples of statistical criteria include calibration, predictive parity, and error rate balance. While these criteria reflect our expectations of a fair decision procedure, tensions have emerged between them.

First, it has been observed that an algorithm can satisfy some of these criteria while failing others, leading to potentially conflicting interpretations of whether the algorithm is fair, as illustrated in the COMPAS case, an algorithm used to predict recidivism. Subsequently, it has been proved that not only does an algorithm not need to satisfy the most prominent fairness criteria at the same time but it cannot satisfy them simultaneously except under special circumstances that are unlikely to occur in real-life settings where our algorithms are deployed. These results, known as the *impossibility results*, elicited various reactions. Some scholars interpreted the impossibility results as the inevitability of bias or the impossibility of complete neutrality towards sensitive traits.

Others, on the other hand, have challenged the relevance or importance of some fairness criteria.

2 The impossibility results and the reactions

2.1 The impossibility results

I begin by introducing the formal framework in which the fairness criteria and the impossibility results will be articulated. The setting for algorithmic fairness is often defines as follows:

1. A profile of sensitive traits $a_i = a_i^1, ..., a_i^n$.
   Sensitive traits may involve gender, race, ethnicity, religion, merital status, age.

2. A profile of permissible traits $x_i = x_i^{n+1}, ..., x_i^m$
   Permissible traits may involve educational background, work experience, technical skills, test scores, credit score.

3. A desired characteristic, $y_i \in \{0, 1\}$
   The desired characteristic can be qualification for a job, aptness for a training program, or a propensity to exhibit some desired behavior (such as repaying credit, avoiding recidivism, or adhering to professional standards).

4. A score $s_i \in [0, 1]$ assessing the likelihood that i has the desired characteristic.

5. A decision, $\delta_i \in \{0, 1\}$
   The decision is the outcome of an evaluative process, such as whether an individual is hired for a job, admitted to a program, granted a loan, awarded a scholarship, selected for a leadership position, or approved for housing, or granted parole.

6. A threshold score $t \in [0, 1]$ used to determine whether individual i is deemed to possess the desired characteristic.

Definition 1 (Calibration)
A score s is well calibrated if it reflects the same likelihood of having the desired characteristic irrespective of group membership. That is, if for all values of s and each pair of groups a_i, a_j:

$$P(Y = 1|S = s, a_i) = P(Y = 1|S = s, a_j)$$

Definition 2 (Predictive Parity)
A score s satisfies *predictive parity* at a threshold t if the likelihood of having the desired characteristic is the same among individuals whose scores are above the threshold, regardless of group membership. That is, if for each pair of groups a_i, a_j:

$$P(Y = 1|S > t, a_i) = P(Y = 1|S > t, a_j)$$

Definition 3 (Error rate balance)
A score s satisfies *error rate balance* at a threshold t if the false positive and false negative error rates are equal accross groups. That is, if each pair of groups a_i, a_j:

$$\begin{aligned} P(S > t|Y = 0, a_i) =& P(S > t|Y = 0, a_j) && (1)\ \{?\} \\ P(S \leq t|Y = 1, a_i) =& P(S \leq t|Y = 1, a_j) && (2)\ \{?\} \end{aligned}$$

Definition 4 (Statistical Parity)
A score s satisfies *statistical parity* at a threshold t if the proportion of individuals whose scores are above the threshold are equal accross the groups. That is, if each pair of groups a_i, a_j:

$$P(S > t|a_i) = P(S > t|a_j)$$

Definition 5 (Balance for the negative and positive classes)
A score s satisfies balance for the negative and positive classes if it satisfies **??** and **??**

?⟨balancenegclass⟩? **Definition 5.1** A score s satisfies *balance for the negative class* the average score assigned to individuals in one group who do not have the desired characteristic must be the same as the average score assigned to individuals in the other group who do not have the desired characteristic. That is, for each pair of groups a_i, a_j:

$$\mathbb{E}[s_i \mid Y = 0, a_i] = \mathbb{E}[s_j \mid Y = 0, a_j]$$

?⟨balanceposclass⟩? **Definition 5.2** A score s satisfies *balance for the positive class* the average score assigned to individuals in one group who have the desired characteristic must be the same as the average score assigned to individuals in the other group who have the desired characteristic. That is, for each pair of groups a_i, a_j:

$$\mathbb{E}[s_i \mid Y = 1, a_i] = \mathbb{E}[s_j \mid Y = 1, a_j]$$

Kleinberg et al. (?) prove that no algorithm can satisfy balance for the negative and positive classes and error rate balance unless either (i) base rates of the desired characteristics are equal across the relevant groups, or (ii) the algorithm is a perfect predictor—i.e. assigns a score of 1 to all those who have the desired characteristic and a score 0 to all those who do not have the desired characteristic. Chouldechova (?) a related result that algorithm cannot satisfy Predictive Parity and Error rate balance at the same time unless the unless base rates are equal or the algorithm is a perfect predictor.

Predictive parity is important because: ....
Error rate balance is important because ....
Calibration is important because ....
Balance in the positive and negative classes is important because

2.2 Responses to the impossibility results

A common reaction to impossibility results is to evaluate which fairness criteria are most important and indispensable. Such efforts can include:

1. Trying to decide between PPV and error rate balance.

?⟨PPVerrorratetradeoff⟩?

2. Trying to weaken PPV and/or error rate balance.

3. Turning to individual metrics– This is also not satisfting. My model will show why individual metrics alone cannot guarentee fairness.

4. Challenging what is considered a necessary condition for fairness by examining algorithms purported to be fair.

?⟨reject⟩?

Before discussing these suggestions, it's important to highlight that the impossibility results in algorithmic fairness literature are challenging because we are under non-ideal conditions. In an ideal scenario, if we had perfect algorithms assigning risk scores of 0 to individuals without the desired characteristic and 1 to those with it, achieving fairness would be straightforward and would not require any trade-offs. However, in practice, our algorithms are far from perfect, leading to trade-offs and difficulties in satisfying multiple fairness criteria simultaneously. Therefore, although the tensions between fairness criteria are indeed seems to be inevitable in real life settings, calling these trade-offs inherent risks obscuring the role that practical limitations play in creating them.

1. Trade-off

One response to the impossibility results is relaxing PPV, positive error rate balance, or negative error rate balance, as suggested in **??**. Chouldechova (?, p. 161) points out that we can tune these signifiers in any of the following ways:

(i) Allow unequal False Negative Rates while maintaining equal Positive Predictive Values and achieving equal False Positive Rates. (ii) Allow unequal False Positive Rates while maintaining equal Positive Predictive Values and achieving equal False Negative Rates. (iii) Allow unequal Positive Predictive Values while maintaining equal False Positive Rates and False Negative Rates.

Chouldechova (?, p. 161) argues that (iii) can be a preferred approach in some cases. This response does not satisfy many people because PPV is still seen as a fairness criteria. SOME REASONS: ...

At any rate, trying to decide between PPV and error rate balance is not a satisfactory way.

2. Weaken

Another approach has been instead of giving up on PPV, weakening it. Some people have suggested base-rate tracking as a weakening of calibration. This is a promisin aproach: however, weakenings of PPV are also subject to impossibility results. (Steward)

3. Turn to individual metrics

Instead, some has turned to individual metrics of fairness giving up on group fairness. This is also not satisfying. I will show that individual metrics are not sufficient for fairness. I will explain this later.

4. Reject the criteria

An example of the efforts characterized by **??** is performed by Brian Hedden. He considers an algorithm which he argues is fair and rejects all the metrics altogether except calibration: arguing that none of them are necessary for a fair algorithm.

The setting Brian Hedden considers is the following: Each individual in a population is randomly assigned a coin with varying biases. Then the individuals are randomly assigned to two rooms, A and B. The goal is to predict whether each person's coin lands heads or tails. Each coin comes labeled with its bias—i.e. its objective chance of landing heads which is a real number in the interval [0,1].

Hedden suggests that the following is a perfectly fair and unbiased predictive algorithm in this setting: For each person, read their coin's bias label and assign them a score equal to the label value. That is, if the bias label says '', assign that person a score of . Set a threshold at 0.5: if > 0.5, predict they are a heads person (positive), and if $x < 0.5$, predict they are a tails person (negative). Assume no coin has a bias of exactly $x = 0.5$, though this assumption can be avoided by making an arbitrary or randomized prediction in such cases.

Hedden stresses that this is algorithm "is not unfair to any people in virtue of their room membership" because its predictions are not sensitive to individuals' room membership. "And the sole feature on which its predictions are based (the labeled bias of the coin) is clearly the relevant one to focus on and is neither a proxy for, nor caused or explained by, room membership." "Indeed, it is not just that the algorithm is in no way unfair to individuals in virtue of their membership in a certain room; there is seemingly no unfairness of any kind anywhere in this situation. Moreover, this algorithm is uniquely optimal; no alternative can be expected to do as well or better at predicting whether individuals are heads people or tails people."

— Hedden argues that calibration cannot be violated.

–will calibration also be satisfied:::: there are many problems with Hedden's setting.

Hedden makes the following observation: all fairness criteria except calibration can be violated simultaneously and even when base rates are equal between the two rooms. (p. 221). Really the striking aspect of Hedden's observation that this fair looking algorithm can violate fairness criteria even when base rates are equal which is seen as the root of the fairness problems (at least the difficulty in satisfying statistical criteria).

See footnote 15 and 35– calibration need to be satisfied purely because he defines it in terms of expectation. but if we define it as .. He assumes that actual frequencies match the expectation.

He considers the following situation (p. 221):

Suppose that the actual relative frequencies match coin biases. Room A contains twelve people coins labeled '0.75' and eight people with coins labeled '0.125'. The former are assigned the score 0.75, therefore, predicted to be heads people; and nine of them turn out to be really heads people. The latter are all assgined a score 0.125, therefore, predicted to be tails people, and but one of them turns out to be a heads person. In room B, there are ten people with coins labeled '0.6' and then with people with coins labeled '0.4'. The former are assigned a score of 0.6, therefore, predicted to be heads people, and six of them turn out to be really heads people. The latter are all assigned risk score 0.4, therefore predicted to be tails people, yet four of them turn out to be heads people.

Hedden takes this to be a counterexample to the necessity of the statistical criteria except calibration. However, the only reason that his setting and algorithm guarantees the satisfaction of calibration is that he defines it as expectation.

1) What this shows is limited because still in the long run, if the assignments are really random, other criteria must be satisfied. 2) Objective chances and coins: the probability distributions are different in different groups. However, this time, this difference of probability distributions drive the result and not the base rate. If we are committed to objective chance distributions, equality of these distributions is more important. 3) Looking at probabilities is a more noisy estimator for room B.

Do we have a good reason to assume that the probability distributions would be different if the base rates were equal? I don't think so. This will

Although I believe his model is not relevant to the context of algorithmic decision-making, his approach is a good starting point. He writes: "One way to determine whether some criterion is a genuine necessary condition on algorithmic fairness is to find a perfectly fair algorithm and see whether it is possible for it to violate that criterion. If so, then the criterion is not in fact a necessary condition on algorithmic fairness." I believe Hedden's proposed fair algorithm is not fair in the context of individuals or at least, it is not relevant. (Importantly, Hedden's algorithm assumes equal base rates) However, there is some merit to his suggestion to start from a perfectly fair algorithm. This is what I will do in this paper. I will consider a fair algorithm and look at which conditions it satisfies. But how a fair algorithm must look requires some theoretical speculation.

3 A fair model

It is common to define fairness which traits are allowed in the decision procedures and which are not. Therefore, just from looking this, we can derive some fairness criteria. One is the following:

Anti-classification: Sensitive traits should not be involved in decision making.

– Some people point out that anti-classification results in more unfairness. But why? This also contributes to perceived tension between fairness criteria? – Notice that the impossibility results were silent about anti-classification so far. They hold regardless how the score was gauged.

– Anti-classification should have two aspects:

Anti-classification about decision: Sensitive traits should not be involved in decision making only the probability estimation must be involved.

Anti-classification about estimation: Sensitive traits should not be involved in estimating the risk scores.

I will not challenge this principle, but try to understand it. Why do we consider involvement of these traits to be unjust? I think it is because we assume that those traits are irrelevant to whether one is qualified or not. We should make our decision on the basis of whether one is qualified or not, not on the basis of one's sensitive traits. This derives from another principle: treat individuals who are known to be similarly qualified similarly!

However, this doesn't explain the contrast between the sensitive traits and the permissible traits. If two people to be equally qualified, then we must treat them in the same way, not only regardless of their sensitive traits but permissible traits as well. In other words, once we have the information about whether an individual has the desired characteristic, anything else is irrelevant. Therefore, if we know the value of

Definition 6 *TREAT EQUALLY QUALIFIED INDIVIDUALS EQUALLY* $P(D|Y,G) = P(D|Y,\neg G)$ *WHEN THE VALUE OF Y IS KNOWN AT THE TIME OF THE DECISION*

The problem is often we don't know whether a person is qualified or not. Therefore, the same equation is often interpreted as equal error rates. This interpretation assumes that treating equally qualified individuals equally as a guiding ideal.

$P(D|Y,G) = P(D|Y,\neg G)$

This condition is often known as error rate balance. But it is not properly called so in this context. If D is interpreted in a certain way. Perfect predictor is required for this but it is not sufficient. We may know that someone is qualified but not hire them on taste-based discrimination.

Nevertheless, we don't know whether one is qualified or not. We just guess it. This is the tricky part.

Already here we can introduce an important distinction: basing decisions on gender vs. basing estimations on gender. Anti-classification fist tells us that given our estimation regarding one's having the desired quality, our decision should be the same across groups.

— Why should we think that our estimation should not depend on the sensitive characteristics? I do believe this is a right principle for the most part. However, it does not directly follow from anti-classification about decision.

When we are trying to estimate whether one is qualified or not, we look at other characteristics—such as education, credit score, etc. Here, too, we think that it is inappropriate to base the estimations on gender–while it is okay to base estimations on some other characteristics. But why? Is it because these are sensitive traits? What makes them sensitive? History of discrimination or even history of oppression is not a sufficient answer. This can make these characteristic particularly salient, but I think that is not that big a problem.

These are problematic because the discrimanation was unfair even back then. What makes discrimination on the basis of education etc acceptable even desired while discrimination on the basis sensitive traits is a principle like the following:

Definition 7 (Anti-essentialism)
There is nothing *inherent* about one's sensitive traits that makes one better or worse qualified. Therefore, any correlation between possessing these sensitive traits and qualification must be driven either by confounding factors or by mediators.

This doesn't explain why these are the sensitive traits. There are many traits that are not relevant to one's qualifications: the favourite ice-cream flavour, zodiac sign, etc.. If it turns out that there is a correlation between these characteristics and qualification, we would look for confounding factors or mediators.

I think this anti-essentialism partly underlies the anti-classification and conditional parity. However, just saying that the sensitive traits should not be involved is not enough to guarantee even minimal fairness. The characteristics that are involved must be relevant to the qualification.

In endorsing conditional demographic parity as a fairness criteria an implicit assumption must be that these are equally good indicators of the deserved quality for both groups. (What is the shark example?) Imagine two types of fish, one turns red when qualified and the other turns green when qualified. It is unfair to ... how to describe the situation.

If there is a test, it should not be depend anything else than whether one is qualified or not– $P(T|Y,G) = P(T|Y,\neg G)$ — the test should treat equals equally: computer based SAT. no one should have an advantage in the test except through being .

(Fair test is sometimes understood as an equally good indication of success. this is exactly what we don't have due to the impossibility results.)

Anti-essentialism, fair test, anti-classification gives us the following Bayes net. Anti-essentialism is not a fairness criteria but it is an assumption we implicitly make in designating certain characteristics as sensitive. Therefore, they might be context dependent. Age can be a sensitive characteristic in job applications or credit decisions, but it is not in medical contexts. (Or pregnancy). I think this intuition partly explain the role of anti-essentialism.

Anti-essentialism tells us we can explain the correlation between Y and G completely through mediating factors. Fair test, if one is – one's test result should not depend anything else other than one is qualified or not.

This model is at least prima facie fair. There is another alternative. Where decision takes E into consideration.

It is helpful to

Individual fairness: Treat similarly qualified individuals similarly.

– This amounts to

$P(D_i|Y = i, G) = P(D_j|Y = j, \neg G)$ iff $i = j$.

However, we don't have direct access to Y. Therefore this becomes the error rate; instead people often mean the following when they say treeat similarly qualified individuals similarly.

Definition 8 (Conditional parity)

An algorithm satisfies **conditional demographic parity** if, for any pair of profiles of sensitive traits, a_i, a_j,

If $x_i = x_2$, then $P(\delta_i|a_i, x_i) = P(\delta_j|a_j, x_j)$

However, interpreting this as treating similarly qualified individuals similarly requires some assumptions: first, $P(Y|X, G) = P(Y|X, \neg G)$ —whatever ???? does it require this

—-

Anti-essentialism

Start with ”treating similarly qualified individuals similarly”. This amounts to ERROR RATE BALANCE. however, the problem is not we cannot observe whether one is really qualified. We have indicators of this. So, ”treating similarly qualified individuals similarly” becomes conditional demographic parity given the fallibility of the decisions.

Definition 9 (Conditional parity)

An algorithm satisfies **conditional demographic parity** if, for any pair of profiles of sensitive traits, a_i, a_j,

If $x_i = x_2$, then $P(\delta_i|a_i, x_i) = P(\delta_j|a_j, x_j)$

Conditional demographic parity is sometimes interpreted as ’treating similarly qualified individuals similarly’. However, for this interpretation to be correct, only if a_i exhausts all the relevant observable qualities an individual have.

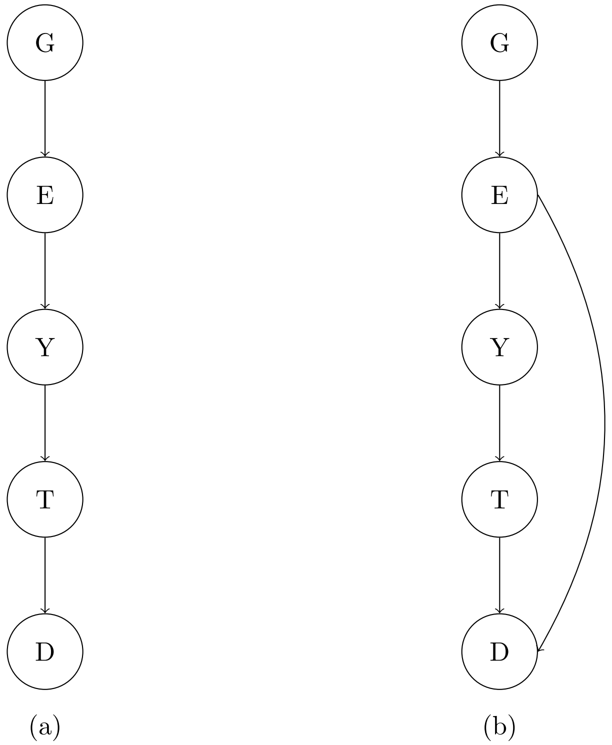
Definition 10 (Fair Test)

An algorithm satisfies **fair test** if, for any pair of profiles of sensitive traits, a_i, a_j,
(If there is any test involved...)

If $y_i = y_2$, then $P(x_i|a_i, y_i) = P(x_j|a_j, y_j)$

Definition 11 (Anti-essentialism)
There is nothing *inherent* about one’s sensitive traits that makes one better or worse qualified. Therefore, any correlation between possessing these sensitive traits and qualification must be driven either by confounding factors or by mediators.

Anti-classification about decision: First, group membership does not influence the decision, ensuring that the procedural fairness condition is satisfied. Anti-classification about estimation: Second, the group membership is not involved in estimating the probability of Y: if E and T. We know this is possible beacause of anti-classification. Third, the test is fair; it is not influenced by anything other than whether the individual is qualified or not.



(a)

(b)

Figure 1: Bayesian Networks. Prima Facie Fair Decision Procedures

?⟨fig:?mylabel⟩?

In this model the relationship between G, E, and Y are external to the algorithm. However, we can assume that there is a mediator between G and Y given anti-essentialism. T can be any test the algorithm is using, such as SAT score. The decision must be either based on . . .

What if there is no test: it is also fine.

Compare these models with unfair procedures.

References

?Chouldechova_2017? Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153-163. doi: 10.1089/big.2016.0047

?Hedden_2021? Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy & Public Affairs*, *49*(2), 209-231. doi: 10.1111/papa.12189

?Kleinberg_2016? Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent tradeoffs in the fair determination of risk scores. *arXiv*. Retrieved from http://arxiv.org/abs/1609.05807 doi: 10.48550/arXiv.1609.05807

?Patty_Penn_2023? Patty, J. W., & Penn, E. M. (2023). Algorithmic fairness and statistical discrimination. *Philosophy Compass*, *18*(1), 1-23. doi: 10.1111/phc3.12891