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 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

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

1. Trying to decide between PPV and error rate balance. – This seems to be a non-ideal way to go, because intuitively both are important and required for fairness.

2. Trying to weaken PPV and/or error rate balance. — This is a promisin aproach: however, weakenings of PPV are also subject to impossibility results.

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.

An example of the efforts characterized by 4 is performed by Brian Hedden. He considers an alogirithm which he argues is fair and rejects all the metrics altogether except calibration: arguing that none of them are necessary for a fair algorithm. 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.

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 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.

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?

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 1: 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 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 1 (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 favouirte 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 2 (Conditional parity)

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

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

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

—-

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 3 (Conditional parity)

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

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

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 4 (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...)

$$\text{If } \; y_i = y_2, \text{ then } \; P(x_i|a_i,y_i) = P(x_j|a_j,y_j)$$

Definition 5 (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.

I suggest this models are fair.

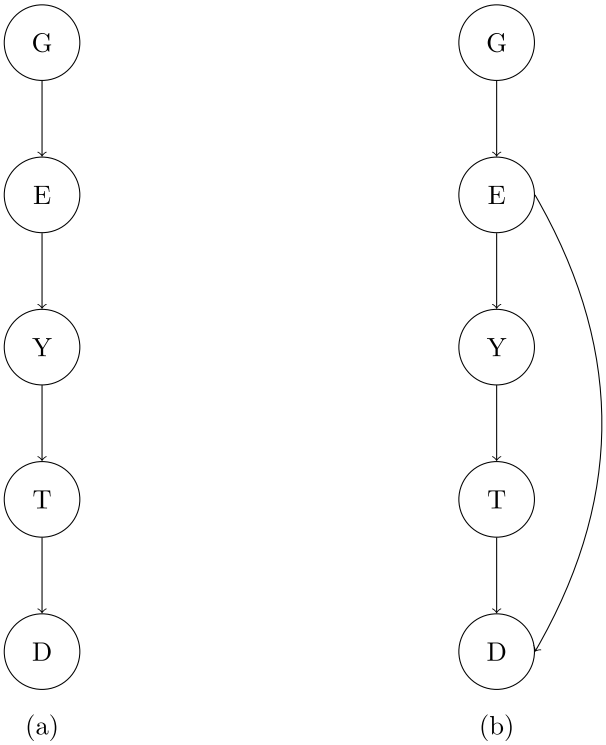

Figure 1: Bayesian Networks. Prima Facie Fair Decision Procedures