# 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 $a_i, a_j$:

$$P(S > t|Y = 0, a_i) = P(S > t|Y = 0, a_j) \tag{1}$$
$$P(S \leq t|Y = 1, a_i) = P(S \leq t|Y = 1, a_j) \tag{2}$$

**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 **??**

**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 $a_i, a_j$:

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

**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.
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.

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. He points out that calibration cannot be violated in this set-up. This is because he defines probabilities as "probabilistic expectations" as opposed to actual relative frequencies (see footnote 15, p. 214). However, he doesn't mean epistemic probability or credence by this at least, in the coin flip case, he suggests that probabilistic expectation can be understood as objective chance. "In others, where no objective chanciness is involved, I suggest that they are best understood as epistemic probabilities, and more specifically as the subjective probabilities that would be assigned by a reasonable individual who is familiar with the workings of the algorithm in question"(see footnote 15, p. 214).

I believe that the main reason we perceive Hedden's algorithm to be fair is the interpretation of probability as objective chances. If the scores assigned by the algorithm did not track objective chances but only represented epistemic probabilities, we would not deem this fair: it would be simply be more informative about one group and less for another. It would be using an estimator that is noisier for the group—without any excuse.

Even in the objective chance reading, the algorithm uses an algorithm that is noisier for a group. Nevertheless, one may argue that this is not the fault of the algorithm but the world is kinder to one group and not to the other. One can still force the noise interpretation: if the bias of the coin is not the only determinant of whether the coin lands heads or tails, by not considering that the algorithm misses more information on one group than on another. But if we interpret the objective chances strictly, there is really nothing the algorithm can do than looking at the coin bias. Brian Hedden's claim holds: there is nothing we can do to improve the algorithm. (notice that if there are other factors, there is a way to improve the algorithm). I will wrap this and assume that Hedden's algorithm is indeed fair in its setting.

Again, what drives the result is inframarginality, as Hedden also points out. Inframarginality is pointed out to undermine the fairness criteria. However, it has not been seen as a source of or a result of unfairness. Here I will argue, unless the inframarginality in the scores tracks a the inframarginality in the objective chances, it is a result of unfairness in the procedure that perpetuates the bias. Now, the question is whether in the real-life settings that the algorithms are deployed there is a difference in the objective chance distribution. This requires answering two questions affirmatively: 1) Are there objective chances in the relevant setting? 2) If yes, are there any differences in the objective chance distributions so that objective probability is a noisier estimator for one group?

Many traits the algorithms are looking for are not chancy, and it is questionable whether any is a chancy event. Even if some of the events are in the future, this doesn't make them chancy. However, I will grant this. Is there any reason to suspect that the chance distribution will be more marginal in one group? If we can rule this out, then we can say...

Noticing the importance of the objective chance distribution also diminishes the strikingness of Hedden's result: he shows that error rate balance and predictive parity are not satisfied even if the base rate is satisfied. However, the chance distribution is significantly different. In this setting objective chances are more important than the ... Thus, we should not read Hedden's result as even if the groups are significantly similar the fairness criteria can be violated.

To take stock: Inframarginality is not an excuse unless objective chances are involved in the relevant setting and the objective chance distibution of groups have inframarginality differences. However, we have no good reasons to think that is the case.

Viganò et al. (?) provide another argument based on the kind of probabilities and the relevancy of Hedden's setup to most of real-life algorithmic settings. They try to avoid any commitment about (ontological) interpretation of probabilities and instead talk about ways of determining probabilities. They distinguish between two ways of computing probabilities: individual-based and group-based. "An h-individual probability is determined with data of the person whom the prediction is about; an h-group probability is determined with data of other people beyond the person whom the prediction is about. These two types of h-probabilities correspond to the two types of probability-based decision-making practices: an h-individual practice is one in which we make a decision about a person by employing the h-individual probabilities attributed to that person; an h-group practice is one in which we take a decision about a person by employing the h-group probabilities attributed to that person"(Viganò et al., 2022, p. 2). They argue that Hedden's argument is valid for individual-based probabilities but does not apply to group-based probabilities about humans. Crucially, "Many practices in data science involve h-group probabilities. This includes the COMPAS example that Hedden treats as paradigmatic."

Their main point is that there is a moral difference to individual- and group-based probabilities. They argue that making decisions on group-based probabilities pro tanto wrongs people because this doesn't treat people as individuals.

"H-individual practices are based on h-individual probabilities, which are estimated only on the basis of what the person did (e.g., being late or punctual, thrifty or spendthrift) or other characteristics (e.g., financial independence, stability of employment, having dependent children) that do not require comparisons with the outcomes of other people to reasonably inform a guess about the future outcomes of that person. Thus, h-individual practices treat a person as an individual. H-group practices are based on h-group probabilities, which are founded on the attribution of the person to a specific group sharing the same or similar aspects with that person (e.g., being a woman, a bank clerk, a member of the same age group, punctual). Thus, h-group practices treat the person as an instance of a more general type, not as an individual." ([Viganò et al., 2022, p. 7](zotero://select/library/items/TPA9HN24))

Thus, Hedden's model is not relevant to real machine learning problems where fairness is a question.

Søgaard et al. (?) contend that Hedden's thought experiment is irrelevant to real machine-learning problems on different grounds. They argue as follows: The performance of classification models is evaluated using hold-out samples—-randomly selected subsets of the dataset excluded from the estimation process. In that, the error of the model is defined as the probability of failing to correctly predict the label for a random data point drawn from the underlying distribution. This principle extends to fairness evaluations: Fairness metrics quantify the expected properties of the underlying distribution, particularly whether the distributions of predictions are consistent across subgroups: "Models are fair if subgroups incur the same loss in the limit" (?, ?, p. 9). In short, fairness in ML is evaluated over distributions, not finite samples, whereas Hedden measures fairness on a specific, small sample of data.

Since in ML held-out samples approximate performance on data distributions, and fairness metrics assess whether subgroup performance converges as more data is sampled independently and at random. To demonstrate flaws in fairness metrics, Hedden would need to show that these metrics consistently score subgroups differently as new data is sampled from the underlying distribution.

Søgaard et al. (?) note that if the classification problem Hedden considers were a genuine ML problem, his algorithm would be evaluated on additional data sampled independently from the underlying distribution.

The fairness of the algorithm would then be assessed based on its loss in the limit, as the sample size approaches infinity. In this scenario, biases in the subgroups would converge to 0.5 as he assumes that base rates are equal across groups and and all fairness metrics would classify the optimal model as fair. This result follows from the Central Limit Theorem, which states that for a population with mean $\mu$ and standard deviation $\sigma$, the distribution of sample means from sufficiently large random samples (with replacement) will approximate a normal distribution with mean $\mu$ and standard deviation $\sigma$.

Any response trying to argue that the future samples from the underlying distribution will be similar to the sample Hedden considers would need to admit that coin distribution or the distribution of people to rooms is non-random. However, as Søgaard et al. (?) write "when coin assignment is non-random, our intuition flips: People in Room 1 are assigned coins from a different probability distribution, which means that the overall distribution of coins is systematically biased – and then Hedden's algorithm becomes unfair for ignoring this systematic bias."

One may still try to defend that Hedden's algorithm is fair: surely, it reproduces the underlyting systematic bias, but it does not have its own bias, so to speak. However, the algorithm uses a measure which is systematically noisier and less informative of one of the groups.

Søgaard et al. conclude that Hedden's algorithm is not relevant to real ML problems. However, an interesting lesson of their observation is that any modification to Hedden's thought experiment that would render it relevant to ML problems and retain our intiuition that the algorithm is fair would satisfy the fairness criteria provided that he assumes the base rates are equal. (We don't have to assume this even, if the distibutions are random, it will be equal.)

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.

Although I believe Hedden's 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. A common idea about fairness is the following:

**Anti-classification:** The decision or evaluation of an algorithm should not depend on individuals's sensitive traits.

This principle is often made more precise as folows: if two individuals with the same permissible traits receive the same decision, $x_i = x_j \Rightarrow \delta_i = \delta_j$ (?, ?, p. 8). I will not use this formulation because this does not guarantee fairness unless other conditions about the permissible traits are satisfied. I will stick to the negative and, temporarily vague, formulation.

Anti-classification is often taken as given; however, I believe it is fruitful to ask why it is taken as a pillar of a fair algorithm. Why do we consider involvement of the sensitive traits to be unfair? I believe, at least one underlying reason is another intuition about fairness: similarly qualified individuals must be treated similarly. The problem is of course, sometimes the decision-makers don't know whether two individuals are similarly qualified because qualification (the desired characteristic)is unobservable. We can then hedge this principle as follows: treat individuals who are known to be similarly qualified similarly.

**Definition 6 (Equal Treatment)**
$P(D|Y, G) = P(D|Y, \neg G)$ when the value of Y is known at the time of 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.

Notice that if an algorithm is perfect predictor, it will be subject to this requirement.

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.

However, when Y is not known, holding on to the same equation might be too strong. Instead, we must hold on to a principle as follows:

**Definition 7 (Equal Treatment, probabilistic)**
$P(D|S = s, G) = P(D|S = s, \neg G)$ when the value of Y is not known at the time of decision

Now, this probabilistic equal treatment condition underlies only one part of anti-classification. Namely, anti-classification about decision:

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

Anti-classification has another part, which can be called anti-classifcation about estimation:

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

Before discussing anti-classification about estimation, it should be noted that equal treatment both in perfect and probabilistic versions, ground more than anti-classification based on sensitive traits: if it is knownn whether an individual is qualified or now, the decision must be based only on this. Nothing else, not only sensitive traits but any other trait the individual may have is irrelevant– whether that trait be one's past training or past experience or horoscope sign. If the qualification is known, they are irrelevant (if they are not, then they must be considred to be a part of the desired characteristic the algorithm is looking for) A similar point can be made about the probabilistic version of this principle.

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

When we move to anti-classification about estimation. This is also considered to be given; but it is not. Why is it unfair to base the estimation about whether one is qualified or not on the sensitive traits? It doesn't follow from anti-classification about decision or the principle that grounds it, the equal treatment. I think it is because we assume that those traits are irrelevant to whether one is qualified or not. 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 8 (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 about estimation. Sometimes this intuition made precise as conditional demogratic parity.

**Definition 9 (Conditional demographic 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, 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. If x is a good indicator for one group but bad for the other, then conditional demographic parity does not reflect fairness. Thus, conditional demographic parity is an indicator of fairness only if another condition is satisfied:

**Definition 10 (Equally good indicator)**
An algorithm satisfies **Equally good indicator** if, for any pair of profiles of sensitive traits, $a_i$, $a_j$,

If $x_i = x_2$, then $P(Y_i|a_i, x_i) = P(Y_j|a_j, x_j)$

However, an implication of the impossibility results is that equally good indicator and conditional demographic parity cannot be satisfied at the same time unless the base rate is equal across groups or the algorithm is a perfect predictor. In the next section, I will show more about the satisfaction conditions of these principles which I will believe instructive. Namely, it is not only that these conditions are simultenously satisfied only under special conditions, but even when they are considered separately they are satisfied under rare conditions.

I argue for a weaker fairness condition: If there is a test, it should not be depend anything else than whether one is qualified or not. The test should treat equally qualified individuals equally. One implication of this condition is the following:

**Definition 11 (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 $P(t_i|a_i, y_i) = P(t_j|a_j, y_j)$

(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 Bayesian networks modelled in **??**. A Bayesian Network represents a set of variables and their conditional independences by organizing them into a *Directed Acyclical Graph* (DAG). A DAG is comprised of a set of nodes and a set of arrows between (some of) the nodes and is constrained by the condition that one cannot form a cycle by following the arrows. The nodes represent the variables and the arrows represent the probabilistic dependence relations between the variables. If two nodes are linked by an arrow, the node at the tail is called the parent node (of the node at the head) and the node at the head is called the child node (of the node at the tail). A node at a tail is called a descendant node of another node if the former can be reached by following arrows starting from the latter. That is, if the the former is a child node of the latter, or a child node of a child node of the latter, and so on.

The arrows in a Bayesian Network carry information about the independence relations between the variables in the network. Parental Markov Condition (PMC) expresses this information:

**Definition 12** *Parental Markov Condition (PMC):* For each variable represented by a node on a Bayesian Network, the variable is probabilistically independent of all variables represented by its non-descendent nodes, conditional on all variables represented by its parent nodes.

Given PMC, we can read off the following information from the Bayesian networks in figure **??**:

**??**a:

$$D \perp Y, E, G|T \tag{1}$$

$$T \perp E, G|Y \tag{2}$$

$$Y \perp G|E \tag{3}$$

**??**b:

$$D \perp Y, G|T, E \tag{4}$$

$$T \perp E, G|Y \tag{5}$$

$$Y \perp G|E \tag{6}$$

1 and 4 ensure that anti-classification about decision is satisfied. They imply that given the values of T and E, the decision is independent of the group membership. 2 and 5 ensure that fair test condition is satisfied, given the value of Y, the test result doesn't depend on anything else, including group membership. 3 and 6 tells us whatever is included in E exhausts the mediators that mediate the relation between G and Y. That there exists a set of traits which fully mediate this relation is an implication of anti-essentialism. This condition is especially important for the second model (**??**b:) because ???? 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. For example, age or pregrancy status can be a sensitive characteristic in job applications or credit decisions, but they may not be a sensitive trait in medical contexts. 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. Not every algorithm relies on a test.

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

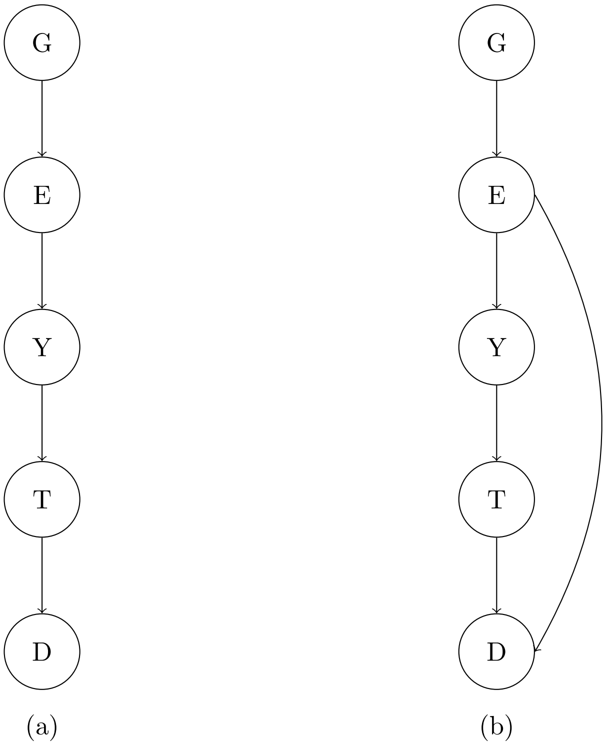
Consider **??**a:

Since this is a Bayes net

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.

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.



(a) (b)

Figure 1: Bayesian Networks. Prima Facie Fair Decision Procedures

Compare these models with unfair procedures.
Results:

1. Error rate balance and PPV cannot be satisfied at the same time unless equal base rate or perfect predictor
2. In a fair algorithm, they are necessarily satisfied if equal base rate or perfect predictor
3. In an unfair algorithm they need not be satisfied even when equal base rate or perfect predictor, even when both are true (You haven't proved this, just a hunch)
4. (Though some unfair algorithm can satisfy them)
5. Conditional on mediators, ...
6. Some unfair algorithms fail to satisfy.
7. However, some unfair algorithms can satisfy them.

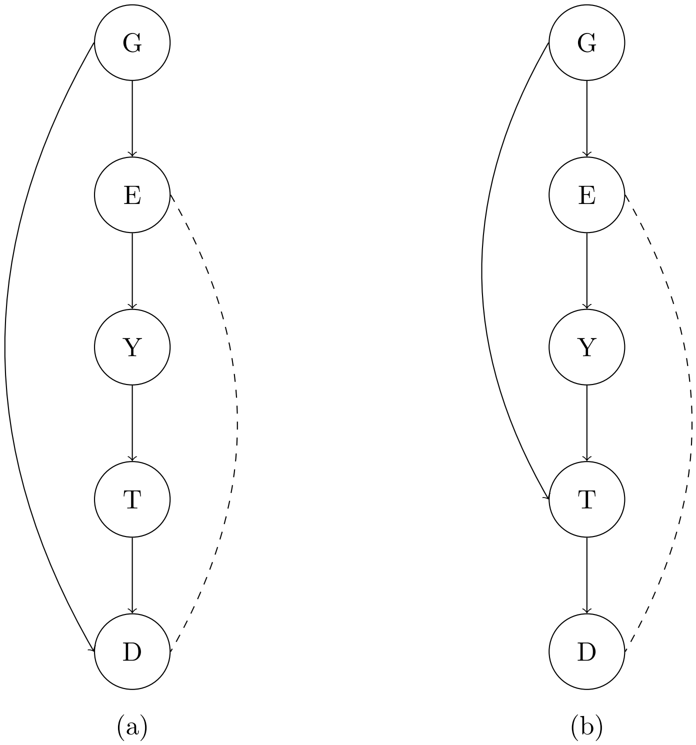


Figure 2: Bayesian Networks. Prima Facie Unfair Decision Procedures

This model is silent about the score. It can be interpreted as P(Y—T) or P(Y—E,T). This model satisfies it.

Full specification of the fair algorithms:

(a)

$$\begin{aligned} P(G) &= g & & \\ P(E|G) &= e & P(E|\neg G) &= f \\ P(Y|E) &= b & P(Y|\neg E) &= c \\ P(T|Y) &= x & P(T|\neg Y) &= y \\ P(D|T) &= \delta & P(D|\neg T) &= \gamma \end{aligned}$$

$$
\begin{aligned}
&P(G) = g \\
&P(E|G) = e \quad P(E|\neg G) = f \\
&P(Y|E) = b \quad P(Y|\neg E) = c \\
&P(T|Y) = x \quad P(T|\neg Y) = y \\
&P(D|T,E) = \delta \quad P(D|\neg T,E) = \gamma \quad P(D|T,\neg E) = \zeta \quad P(D|\neg T,\neg E) = \epsilon
\end{aligned}
$$

Predictive parity:

$$
\begin{aligned}
P(Y|D,G) &= \frac{P(Y,D,G)}{P(D,G)} \\
&= \frac{\sum_{T,E} P(Y,D,G,T,E)}{\sum_{T,E,Y} P(Y,D,G,T,E)} \\
&= \frac{\begin{array}{c} P(D|T)P(T|Y)P(Y|E)P(E|G)P(G) + P(D|T)P(T|Y)P(Y|\neg E)P(\neg E|G)P(G) \\ + P(D|\neg T)P(\neg T|Y)P(Y|E)P(E|G)P(G) + P(D|\neg T)P(\neg T|Y)P(Y|\neg E)P(\neg E|G)P(G) \end{array}}{\begin{array}{l} P(D|T)P(T|Y)P(Y|E)P(E|G)P(G) + P(D|T)P(T|Y)P(Y|\neg E)P(\neg E|G)P(G) \\ + P(D|\neg T)P(\neg T|Y)P(Y|E)P(E|G)P(G) + P(D|\neg T)P(\neg T|Y)P(Y|\neg E)P(\neg E|G)P(G) \\ + P(D|T)P(T|\neg Y)P(\neg Y|E)P(E|G)P(G) + P(D|T)P(T|\neg Y)P(\neg Y|\neg E)P(\neg E|G)P(G) \\ + P(D|\neg T)P(\neg T|\neg Y)P(\neg Y|E)P(E|G)P(G) + P(D|\neg T)P(\neg T|\neg Y)P(\neg Y|\neg E)P(\neg E|G)P(G) \end{array}} \\
&= \frac{\delta xbeg + \delta xc(1-e)g + \gamma(1-x)beg + \gamma(1-x)c(1-e)g}{\delta xbeg + \delta xc(1-e)g + \gamma(1-x)beg + \gamma(1-x)c(1-e)g +} \\
\delta y(1-b)eg &+ \delta y(1-c)(1-e)g + \gamma(1-y)(1-b)eg + \gamma(1-y)(1-c)(1-e)g \\
&= \frac{g(\delta xbe + \delta xc(1-e) + \gamma(1-x)be + \gamma(1-x)c(1-e))}{g(\delta xbe + \delta xc(1-e) + \gamma(1-x)be + \gamma(1-x)c(1-e) +} \\
\delta y(1-b)e &+ \delta y(1-c)(1-e) + \gamma(1-y)(1-b)e + \gamma(1-y)(1-c)(1-e)) \\
&= \frac{\delta xbe + \delta xc(1-e) + \gamma(1-x)be + \gamma(1-x)c(1-e)}{\delta xbe + \delta xc(1-e) + \gamma(1-x)be + \gamma(1-x)c(1-e) +} \\
\delta y(1-b)e &+ \delta y(1-c)(1-e) + \gamma(1-y)(1-b)e + \gamma(1-y)(1-c)(1-e) \\
&= \frac{\delta xbe + \delta xc - \delta xce + \gamma be - \gamma xbe + \gamma c - \gamma xc - \gamma ce + \gamma xce}{\delta xbe + \delta xc - \delta xce + \gamma be - \gamma xbe + \gamma c - \gamma xc - \gamma ce + \gamma xce +} \\
\delta ye &- \delta ybe + \delta y - \delta yc - \delta ye + \delta yce + \gamma e - \gamma ye - \gamma be + \gamma ybe + \gamma - \gamma y - \gamma c + \gamma yc - \\
\gamma e &+ \gamma ye + \gamma ce - \gamma yce \\
&= \frac{e(\delta xb - \delta xc + \gamma b - \gamma xb - \gamma c + \gamma xc) + c(\delta x + \gamma - \gamma x)}{\delta xbe + \delta xc - \delta xce + \cancel{\gamma b}e - \gamma xbe + \cancel{\gamma c} - \gamma xc - \cancel{\gamma ce} + \gamma xce + \cancel{\delta y}e - \delta ybe +} \\
\delta y &- \delta yc - \cancel{\delta y}e + \delta yce + \cancel{\gamma e} - \cancel{\gamma ye} - \cancel{\gamma b}e + \gamma ybe + \gamma - \gamma y - \cancel{\gamma c} - + \gamma yc \\
&- \cancel{\gamma e} + \cancel{\gamma ye} + \cancel{\gamma ce} - \gamma yce \\
&= \frac{e(\delta xb - \delta xc + \gamma b - \gamma xb - \gamma c + \gamma xc) + c(\delta x + \gamma - \gamma x)}{\delta xbe + \delta xc - \delta xce - \gamma xbe - \gamma xc + \gamma xce - \delta ybe +} \\
\delta y &- \delta yc + \delta yce + + \gamma ybe + \gamma - \gamma y + \gamma yc - \gamma yce \\
&= \frac{e(\delta xb - \delta xc + \gamma b - \gamma xb - \gamma c + \gamma xc) + c(\delta x + \gamma - \gamma x)}{e(xb(\delta-\gamma) - xc(\delta-\gamma) - yb(\delta-\gamma) + yc(\delta-\gamma)) + xc(\delta-\gamma) - yc(\delta-\gamma) + y(\delta-\gamma) + \gamma} \\
&= \frac{e(xb(\delta-\gamma) - xc(\delta-\gamma) + \gamma(b-c)) + c(x(\delta-\gamma) + \gamma)}{e(xb(\delta-\gamma) - xc(\delta-\gamma) - yb(\delta-\gamma) + yc(\delta-\gamma)) + xc(\delta-\gamma) - yc(\delta-\gamma) + y(\delta-\gamma) + \gamma}
\end{aligned}
$$

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

$$=\frac{\sum_{T,E}P(Y,D,\neg G,T,E)}{\sum_{T,E,Y}P(Y,D,\neg G,T,E)}$$

$$=\frac{\begin{array}{l}P(D|T)P(T|Y)P(Y|E)P(E|\neg G)P(\neg G)+P(D|T)P(T|Y)P(Y|\neg E)P(\neg E|\neg G)P(\neg G)\\ \quad +P(D|\neg T)P(\neg T|Y)P(Y|E)P(E|\neg G)P(\neg G)+P(D|\neg T)P(\neg T|Y)P(Y|\neg E)P(\neg E|\neg G)P(\neg G)\end{array}}{\begin{array}{l}P(D|T)P(T|Y)P(Y|E)P(E|\neg G)P(\neg G)+P(D|T)P(T|Y)P(Y|\neg E)P(\neg E|\neg G)P(\neg G)\\ +P(D|\neg T)P(\neg T|Y)P(Y|E)P(E|\neg G)P(\neg G)+P(D|\neg T)P(\neg T|Y)P(Y|\neg E)P(\neg E|\neg G)P(\neg G)\\ +P(D|T)P(T|\neg Y)P(\neg Y|E)P(E|\neg G)P(\neg G)+P(D|T)P(T|\neg Y)P(\neg Y|\neg E)P(\neg E|\neg G)P(\neg G)\\ +P(D|\neg T)P(\neg T|\neg Y)P(\neg Y|E)P(E|\neg G)P(\neg G)+P(D|\neg T)P(\neg T|\neg Y)P(\neg Y|\neg E)P(\neg E|\neg G)P(\neg G)\end{array}}$$

$$=\frac{\delta xbf(1-g)+\delta xc(1-f)(1-g)+\gamma(1-x)bf(1-g)+\gamma(1-x)c(1-f)(1-g)}{\delta xbf(1-g)+\delta xc(1-f)(1-g)+\gamma(1-x)bf(1-g)+\gamma(1-x)c(1-f)(1-g)+}$$

$$\delta y(1-b)f(1-g)+\delta y(1-c)(1-f)(1-g)+\gamma(1-y)(1-b)f(1-g)+\gamma(1-y)(1-c)(1-f)(1-g)$$

$$=\frac{(1-g)(\delta xbf+\delta xc(1-f)+\gamma(1-x)bf+\gamma(1-x)c(1-f))}{(1-g)(\delta xbf+\delta xc(1-f)+\gamma(1-x)bf+\gamma(1-x)c(1-f)+}$$

$$\delta y(1-b)f+\delta y(1-c)(1-f)+\gamma(1-y)(1-b)f+\gamma(1-y)(1-c)(1-f))$$

$$=\frac{\delta xbf+\delta xc(1-f)+\gamma(1-x)bf+\gamma(1-x)c(1-f)}{\delta xbf+\delta xc(1-f)+\gamma(1-x)bf+\gamma(1-x)c(1-f)+}$$

$$\delta y(1-b)f+\delta y(1-c)(1-f)+\gamma(1-y)(1-b)f+\gamma(1-y)(1-c)(1-f)$$

$$=\frac{\delta xbf+\delta xc-\delta xcf+\gamma bf-\gamma xbf+\gamma c-\gamma xc-\gamma cf+\gamma xcf}{\delta xbf+\delta xc-\delta xcf+\gamma bf-\gamma xbf+\gamma c-\gamma xc-\gamma cf+\gamma xcf+}$$

$$\delta yf-\delta ybf+\delta y-\delta yc-\delta yf+\delta ycf+\gamma f-\gamma yf-\gamma bf+\gamma ybf+\gamma-\gamma y-\gamma c+\gamma yc-$$

$$\gamma f+\gamma yf+\gamma cf-\gamma ycf$$

$$=\frac{f(\delta xb-\delta xc+\gamma b-\gamma xb-\gamma c+\gamma xc)+c(\delta x+\gamma-\gamma x)}{\delta xbf+\delta xc-\delta xcf+\cancel{\gamma bf}-\gamma xbf+\cancel{\gamma c}-\gamma xc-\cancel{\gamma cf}+\gamma xcf+\cancel{\delta yf}-\delta ybf+}$$

$$\delta y-\delta yc-\cancel{\delta yf}+\delta ycf+\cancel{\gamma f}-\cancel{\gamma yf}-\cancel{\gamma bf}+\gamma ybf+\gamma-\gamma y-\cancel{\gamma c}-+\gamma yc$$

$$-\cancel{\gamma f}+\cancel{\gamma yf}+\cancel{\gamma cf}-\gamma ycf$$

$$=\frac{f(\delta xb-\delta xc+\gamma b-\gamma xb-\gamma c+\gamma xc)+c(\delta x+\gamma-\gamma x)}{\delta xbf+\delta xc-\delta xcf-\gamma xbf-\gamma xc+\gamma xcf-\delta ybf+}$$

$$\delta y-\delta yc+\delta ycf++\gamma ybf+\gamma-\gamma y+\gamma yc-\gamma ycf$$

$$=\frac{f(\delta xb-\delta xc+\gamma b-\gamma xb-\gamma c+\gamma xc)+c(\delta x+\gamma-\gamma x)}{f(xb(\delta-\gamma)-xc(\delta-\gamma)-yb(\delta-\gamma)+yc(\delta-\gamma))+xc(\delta-\gamma)-yc(\delta-\gamma)+y(\delta-\gamma)+\gamma}$$

$$=\frac{f(xb(\delta-\gamma)-xc(\delta-\gamma)+\gamma(b-c))+c(x(\delta-\gamma)+\gamma)}{f(xb(\delta-\gamma)-xc(\delta-\gamma)-yb(\delta-\gamma)+yc(\delta-\gamma))+xc(\delta-\gamma)-yc(\delta-\gamma)+y(\delta-\gamma)+\gamma}$$

Predictive parity:

$$P(Y|D,G) = \frac{e(xb(\delta - \gamma) - xc(\delta - \gamma) + \gamma(b - c)) + c(x(\delta - \gamma) + \gamma)}{e(xb(\delta - \gamma) - xc(\delta - \gamma) - yb(\delta - \gamma) + yc(\delta - \gamma)) + xc(\delta - \gamma) - yc(\delta - \gamma) + y(\delta - \gamma) + \gamma}$$

$$P(Y|D,\neg G) = \frac{f(xb(\delta - \gamma) - xc(\delta - \gamma) + \gamma(b - c)) + c(x(\delta - \gamma) + \gamma)}{f(xb(\delta - \gamma) - xc(\delta - \gamma) - yb(\delta - \gamma) + yc(\delta - \gamma)) + xc(\delta - \gamma) - yc(\delta - \gamma) + y(\delta - \gamma) + \gamma}$$

Error rate balance:

# References

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

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

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