PHIL 408Q/PHPE 308D
Fairness

Eric Pacuit, University of Maryland

April 16, 2024

John W. Patty and Elizabeth Maggie Penn (2022). *Algorithmic Fairness and Statistical Discrimination*. Philosophy Compass.

Algorithmic Fairness: Algorithmic fairness (AF) is a new term describing the study of how to evaluate rule-based procedures for making decisions about diverse individuals. At the heart of this study is the presumption that certain ways of discriminating between two or more individuals are undesirable (i.e., "unfair"), whereas others are less suspect, or even desirable (i.e., "permissible").

Algorithmic Fairness: Algorithmic fairness (AF) is a new term describing the study of how to evaluate rule-based procedures for making decisions about diverse individuals. At the heart of this study is the presumption that certain ways of discriminating between two or more individuals are undesirable (i.e., "unfair"), whereas others are less suspect, or even desirable (i.e., "permissible").

Statistical Discrimination: The literature on statistical discrimination (SD) is more established than that on AF. Rather than measuring and classifying disparities in algorithmic performance across groups, this literature squarely aims to identify the root causes of discrimination, and to disentangle disparate outcomes due to discrimination (i.e., disparate treatment) from those due to exogenous disparities across groups.

Hiring

Each applicant has a single, unobserved characteristic that is of interest to the decision-maker (e.g., is the individual "qualified" for the job or not).

For any applicant, the "hiring algorithm" (which might "represent a strategic employer" or not) makes a binary choice (e.g. to hire or not).

Hiring a qualified individual or not hiring an unqualified individual are each considered a success, while hiring an unqualified applicant or not hiring a qualified applicant are each considered failures of the algorithm.

Suppose that E is an employer and $N = \{1, 2, \ldots, n\}$ is a pool of applicants.

Each applicant $i \in N$ is described by the following:

Each applicant $i \in N$ is described by the following:

1. A profile of **permissible traits** $x_i = (x_i^1, \ldots, x_i^m)$

   Examples: Education, technical skills, test scores, credit history.

Each applicant $i \in N$ is described by the following:

1. A profile of **permissible traits** $x_i = (x_i^1, \ldots, x_i^m)$

   Examples: Education, technical skills, test scores, credit history.

2. A profile of **sensitive traits** $a_i = (a_i^1, \ldots, a_i^k)$.

   Examples: Gender, race, ethnicity, marital status.

Each applicant $i \in N$ is described by the following:

1. A profile of **permissible traits** $x_i = (x_i^1, \ldots, x_i^m)$

   Examples: Education, technical skills, test scores, credit history.

2. A profile of **sensitive traits** $a_i = (a_i^1, \ldots, a_i^k)$.

   Examples: Gender, race, ethnicity, marital status.

3. An **outcome** $y_i \in \{0, 1\}$

   Examples: Qualification for the job, profitability of investment, efficacy of treatment.

Each applicant $i \in N$ is described by the following:

1. A profile of **permissible traits** $x_i = (x_i^1, \ldots, x_i^m)$

   Examples: Education, technical skills, test scores, credit history.

2. A profile of **sensitive traits** $a_i = (a_i^1, \ldots, a_i^k)$.

   Examples: Gender, race, ethnicity, marital status.

3. An **outcome** $y_i \in \{0, 1\}$

   Examples: Qualification for the job, profitability of investment, efficacy of treatment.

4. A **decision** $\delta_i \in \{0, 1\}$

   Examples: Did i get the job? Did i get admitted? Did i get the loan?

	Positive ($\delta = 1$)	Positive ($\delta = 0$)	
Positive ($y = 1$)	True Positive (TP)	False Negative (FN)	TPR: $\frac{TP}{TP+FN}$
Negative ($y = 0$)	False Positive (FP)	False Negative (TN)	TNR: $\frac{FP}{FP+TN}$
	PPV: $\frac{TP}{TP+FP}$	NPV: $\frac{TN}{TN+FN}$	

Anti-Classification

Anti-classification: Sensitive traits are not directly used to make decisions.

An algorithm satisfies anti-classification if two individuals with the same permissible traits receive the same decision, or:

$$x_i = x_j \Rightarrow \delta_i = \delta_j$$

Anti-Classification

Anti-classification: Sensitive traits are not directly used to make decisions.

An algorithm satisfies anti-classification if two individuals with the same permissible traits receive the same decision, or:

$$x_i = x_j \Rightarrow \delta_i = \delta_j$$

- Anti-classification restricts the information that decisions can be responsive to.
- It is clearly associated with process: what factors can directly affect the algorithm's decision for any given individual?
- It is also trivially satisfiable. For example, anti-classification is satisfied simply by having the algorithm assign every individual the same decision ($\delta_i = \delta_j$ for all i, j)

Compas Data

Overall population (18,293 defendants)

	high risk ($\delta = 1$)	nonhigh risk ($\delta = 0$)
actually recidivist ($y = 1$)	2921	5489
actually non-recidivist ($y = 0$)	1693	8190

True Positive Rate: $\frac{2921}{2921+5489} \approx 0.347$

False Positive Rate: $\frac{1693}{1693+8190} \approx 0.171$

$a = 1$: sub-population (9779 black defendants)

	high risk ($\delta = 1$)	nonhigh risk ($\delta = 0$)
actually recidivist ($y = 1$)	2174	2902
actually non-recidivist ($y = 0$)	1226	3477

$a = 0$: sub-population (8514 nonblack defendants)

	high risk ($\delta = 1$)	nonhigh risk ($\delta = 0$)
actually recidivist ($y = 1$)	747	2587
actually non-recidivist ($y = 0$)	467	4713

Predictive Parity

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

Predictive Parity

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$PPV = \frac{2174}{2174+1226} \approx 0.639$

Predictive Parity

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$PPV = \frac{2174}{2174+1226} \approx 0.639$

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$PPV = \frac{747}{747+467} \approx 0.615$

Predictive Parity

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$PPV = \frac{2174}{2174+1226} \approx 0.639$

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$PPV = \frac{747}{747+467} \approx 0.615$

$$0.639 \approx 0.615$$

Predictive parity captures the idea that, conditional on the decision δ, individuals with different sensitive traits should be equally likely to have the same outcome.

Error Rate Balance

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

Error Rate Balance

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$$TPR = \frac{2174}{2174+2902} \approx 0.428$$
$$FPR = \frac{1226}{1226+3477} \approx 0.261$$

Error Rate Balance

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$$TPR = \frac{2174}{2174+2902} \approx 0.428$$
$$FPR = \frac{1226}{1226+3477} \approx 0.261$$

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$$TPR = \frac{747}{747+2587} \approx 0.224$$
$$FPR = \frac{467}{467+4713} \approx 0.090$$

Error Rate Balance

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$$TPR = \frac{2174}{2174+2902} \approx 0.428$$
$$FPR = \frac{1226}{1226+3477} \approx 0.261$$

$$TPR = \frac{747}{747+2587} \approx 0.224$$
$$FPR = \frac{467}{467+4713} \approx 0.090$$

$$0.428 \not\approx 0.224 \text{ and } 0.261 \not\approx 0.090$$

Error rate balance requires that individuals differing only with respect to sensitive traits are equally likely to be misclassified by the algorithm.

Demographic Parity (Statistical Parity)

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

Demographic Parity (Statistical Parity)

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$$\frac{2174+1226}{2174+2902+1226+3477} \approx 0.348$$

Demographic Parity (Statistical Parity)

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$\frac{2174+1226}{2174+2902+1226+3477} \approx 0.348$

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$\frac{747+467}{747+2587+467+4713} \approx 0.143$

Demographic Parity (Statistical Parity)

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$\frac{2174+1226}{2174+2902+1226+3477} \approx 0.348$ $\frac{747+467}{747+2587+467+4713} \approx 0.143$

$$0.348 \not\approx 0.143$$

Demographic parity (sometimes referred to as statistical parity or group fairness) is a widely employed fairness criterion. Substantively, demographic parity is satisfied when sensitive traits do not affect the distribution of decisions for a randomly drawn individual.

Base Rates

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

Base Rates

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$\frac{2174+2902}{2174+2902+1226+3477} \approx 0.519$

Base Rates

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$\frac{2174+2902}{2174+2902+1226+3477} \approx 0.519$

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$\frac{747+2587}{747+2587+467+4713} \approx 0.392$

Base Rates

$a = 1$

	$\delta = 1$	$\delta = 0$
$y = 1$	2174	2902
$y = 0$	1226	3477

$a = 0$

	$\delta = 1$	$\delta = 0$
$y = 1$	747	2587
$y = 0$	467	4713

$\frac{2174+2902}{2174+2902+1226+3477} \approx 0.519$

$\frac{747+2587}{747+2587+467+4713} \approx 0.392$

$$0.519 \not\approx 0.392$$

The base rates of recidivism are not equal.

So, there is a conflict between different notions of fairness when analyzing the COMPAS algorithm.

Game-Theoretic Models of Discrimination

- When the workers' sensitive trait is observed by the employer at the time of making the hiring decision, individuals with different sensitive traits may be treated by the employer differently in the sense that the hiring rule for one group is different from the hiring rule applied to a different group.

Game-Theoretic Models of Discrimination

- When the workers' sensitive trait is observed by the employer at the time of making the hiring decision, individuals with different sensitive traits may be treated by the employer differently in the sense that the hiring rule for one group is different from the hiring rule applied to a different group.

- This, in turn, leads to each worker's incentive to invest in obtaining qualification endogenously depending on the worker's sensitive trait.

Game-Theoretic Models of Discrimination

- When the workers' sensitive trait is observed by the employer at the time of making the hiring decision, individuals with different sensitive traits may be treated by the employer differently in the sense that the hiring rule for one group is different from the hiring rule applied to a different group.

- This, in turn, leads to each worker's incentive to invest in obtaining qualification endogenously depending on the worker's sensitive trait.

- Accordingly, discriminatory behavior by the employer may emerge as a result of the equilibrium played by the employer and worker depending on the worker's sensitive trait (in game theoretic terms, this is referred to as equilibrium selection).

For example, it can be the case that the employer believes that women invest in qualification with some positive probability, but that men do not. In this case, the employer may (correctly) be willing to hire women whose test scores are high enough but (correctly) never hire a male applicant regardless of his or her test score.

This type of discriminatory equilibrium can emerge even if men and women are otherwise identical.

Brian Hedden (2021). *On statistical criteria of algorithmic fairness*. Philosophy & Public Affairs, 49(2), pp. 209 - 231.

Predictive Algorithms

Algorithms such as COMPAS are *predictive algorithms*: They focus on making *predictions* rather than making *decisions*.

Predictive Algorithms

Algorithms such as COMPAS are *predictive algorithms*: They focus on making *predictions* rather than making *decisions*.

Given an input of *features*, typically called a *feature vector*, output a *binary prediction* or a *risk score*:

- The binary prediction (e.g., 0 or 1) classifies individuals as either 'positive' (label 1) or 'negative' (label 0);
- The risk score should be thought of as the probability that the individual falls in the 'positive class'.

Predictions vs. Decisions

A predictive algorithm might be perfectly fair, even though its predictions are put to subtly unfair or even blatantly nefarious uses.

Moreover, a single predictive algorithm might be put to multiple uses, some benign and some not, or it might not feed into any decisions at all, being used instead just to satisfy one's curiosity.

Fairness

"I want to focus not on whether an algorithm is unfair to individuals, or whether it is unfair to groups. Rather, I want to focus on whether it is unfair to individuals *in virtue of their membership in a certain group*."

Fairness

How does this notion of fairness differ from the others?

Fairness

How does this notion of fairness differ from the others?

- One can be unfair to an individual without being unfair to them in virtue of their group membership.

Fairness

How does this notion of fairness differ from the others?

- One can be unfair to an individual without being unfair to them in virtue of their group membership.
- It is not obvious that fairness is owed to groups, as opposed to individuals.

Fairness

How does this notion of fairness differ from the others?

- One can be unfair to an individual without being unfair to them in virtue of their group membership.
- It is not obvious that fairness is owed to groups, as opposed to individuals.
- Granting the notion of unfairness to groups, one can perhaps be unfair to an individual in virtue of their membership in a certain group without being unfair to that group itself, for instance if one treats a single individual worse because of their race or gender but at the same time takes other actions that are to the net benefit of that group.

Fairness Criterion

Some fairness criterion involve studying the internal workings of the algorithm. E.g., the algorithm cannot be based on certain features.

Fairness Criterion

Some fairness criterion involve studying the internal workings of the algorithm. E.g., the algorithm cannot be based on certain features.

Statistical Criteria of Fairness: Criteria that require that certain relations between predictions and actuality be the same for each of the groups in question.

The criteria can be evaluated without actually looking at the inner workings of the algorithm, which may be proprietary or otherwise opaque. Instead, we just have look at the results—what the algorithm predicted and what actually happened.

Fairness for Risk Scores

Calibration Within Groups: For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.

Fairness for Risk Scores

Calibration Within Groups: For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.

The idea is that fairness requires a given risk score to "mean the same thing" for each relevant group. We want the assignment of a given risk score to have the same evidential value, regardless of the group to which the individual belongs.

Fairness for Binary Predictions

Equal Positive Predictive Value: The (expected) percentage of individuals predicted to be positive who are actually positive is the same for each relevant group.

Equal Negative Predictive Value: The (expected) percentage of individuals predicted to be negative who are actually negative is the same for each relevant group.

Fairness for Binary Predictions

Equal Positive Predictive Value: The (expected) percentage of individuals predicted to be positive who are actually positive is the same for each relevant group.

Equal Negative Predictive Value: The (expected) percentage of individuals predicted to be negative who are actually negative is the same for each relevant group.

The idea is that fairness requires a prediction of positive to mean the same thing, or to have the same evidential value, regardless of the group to which the individual belongs (similarly for a prediction of negative).

Fairness for Binary Predictions

Equal False-Positive Rates: The (expected) percentage of actually negative individuals who are falsely predicted to be positive is the same for each relevant group.

Equal False-Negative Rates: The (expected) percentage of actually positive individuals who are falsely predicted to be negative is the same for each relevant group.

Fairness for Binary Predictions

Equal False-Positive Rates: The (expected) percentage of actually negative individuals who are falsely predicted to be positive is the same for each relevant group.

Equal False-Negative Rates: The (expected) percentage of actually positive individuals who are falsely predicted to be negative is the same for each relevant group.

The idea is that fairness requires individuals from different groups who exhibit the same behavior to, on balance, be treated the same by the algorithm in terms of whether they are predicted to be positive or negative. It would be unfair, for instance, if individuals from one group who are actually negative tended to be predicted to be positive at higher rates than actually negative members of the other group.

Fairness for Risk Scores

Balance for the Positive Class: The (expected) average risk score assigned to those individuals who are actually positive is the same for each relevant group.

Balance for the Negative Class: The (expected) average risk score assigned to those individuals who are actually negative is the same for each relevant group.

These are generalizations of the previous two conditions from the case of binary predictions to the case of risk scores, and are motivated in the same way.

Fairness for Risk Scores

Equal Ratios of False-Positive Rate to False-Negative Rate: The (expected) ratio of the false-positive rate to the false-negative rate is the same for each relevant group.

Equal Overall Error Rates: The (expectation of) the number of false positives and false negatives, divided by the number of individuals, is the same for each relevant group.

Fairness for Risk Scores

Equal Ratios of False-Positive Rate to False-Negative Rate: The (expected) ratio of the false-positive rate to the false-negative rate is the same for each relevant group.

Equal Overall Error Rates: The (expectation of) the number of false positives and false negatives, divided by the number of individuals, is the same for each relevant group.

The idea is that fairness requires assigning equal relative weights to the two main error types, false positives and false negatives, for the various groups. It would be unfair, for instance, if the algorithm tended to err on the side of caution for one group while tending to do the reverse for the other group.

Equal Overall Error Rates incorporates the thought that it would be unfair if an algorithm were simply less accurate for one group than for another.

Fairness for Risk Scores

Statistical Parity: The (expected) percentage of individuals predicted to be positive is the same for each relevant group.

Fairness for Risk Scores

Statistical Parity: The (expected) percentage of individuals predicted to be positive is the same for each relevant group.

The idea is that the percentage of individuals predicted to be positive be the same for each relevant group.

However, this criteria is in fact widely rejected, because it is insensitive to differences in base rates (ratios of actual positives to actual negatives) across groups. Indeed, when base rates differ across groups, this criterion will be violated by an omniscient algorithm which perfectly predicts people's behavior. But a perfect algorithm would, presumably, not be unfair simply in virtue of differing base rates.

Fairness for Risk Scores

Equal Ratios of Predicted Positives to Actual Positives: The (expectation of) the number of individuals predicted to be positive, divided by the number of individuals who are actually positive, is the same for each relevant group.

Fairness for Risk Scores

Equal Ratios of Predicted Positives to Actual Positives: The (expectation of) the number of individuals predicted to be positive, divided by the number of individuals who are actually positive, is the same for each relevant group.

This improves on the previous previous criterion. When base rates differ, this requires that differences in base rates yield corresponding differences in the rates at which individuals are predicted to be positive.