# Algorithmic Fairness: A Primer on Impossibility Theorems

Tina Eliassi-Rad & Branden Fitelson

http://fitelson.org/primer.pdf



Nowadays, the use of machine learning algorithms to classify (or make predictions about) human beings is ubiquitous.

Often, these algorithms make predictions/classifications that are *biased* (or "unfair") with respect to certain protected/sensitive characteristics (*e.g.*, gender, ethnicity, sexual orientation) [8].

For instance, the COMPAS recidivism prediction algorithm has exhibited various kinds of bias (or unfairness) with respect to the protected attribute of race [1]. More on this case, below.

There has been a lot of recent work on "fairness in machine learning," some of which is devoted to developing and applying statistical *fairness measures* to address these problems.

Next, we will describe some of these fairness measures. Then, we will discuss some impossibility theorems, which reveal that not all of these notions of fairness can be satisfied simultaneously.

Time permitting, we will briefly demo our *Mathematica* [5] notebook for exploring these impossibility theorems [4].





Suppose we are evaluating an algorithm which aims to predict whether a person (in some population) will have some target property (on the basis of some known characteristics).

Let $T$ denote the proposition that the person (*in fact*) has the target property, and let $\hat{T}$ denote the proposition that *the algorithm predicts that* person has the target property.

Finally, let $P$ denote the proposition that the person has some protected/sensitive characteristic or property.

In the COMPAS example, $T$ states that the person will (likely) recidivate, and $P$ states that the person in question is black.

All of the fairness measures we will discuss involve various probabilistic relations among the three propositions $T$, $\hat{T}$, and $P$.

Specifically, all of the fairness measures will involve relations of *probabilistic independence* among the propositions $T$, $\hat{T}$, and $P$.

Next: two key notions of probabilistic independence.





**Definition**. Propositions $X$ and $Y$ are (unconditionally) *independent* (abbreviated $X \perp\!\!\!\perp Y$) just in case

$$\Pr(X \mid Y) = \Pr(X \mid \neg Y).$$

In words, $X$ and $Y$ are (unconditionally) *independent* ($X \perp\!\!\!\perp Y$) iff $X$ has the same probability on the supposition of $Y$ as it does on the supposition of $\neg Y$. Note: the relation $\perp\!\!\!\perp$ is *symmetric*.

**Definition**. Propositions $X$ and $Y$ are *conditionally independent, given proposition* $Z$ (abbreviated $X \perp\!\!\!\perp Y \mid Z$) just in case

$$\Pr(X \mid Y \,\&\, Z) = \Pr(X \mid \neg Y \,\&\, Z).$$

In words, $X$ and $Y$ are *conditionally independent, given* $Z$ ($X \perp\!\!\!\perp Y \mid Z$) iff *supposing* $Z$ renders $X$ and $Y$ independent.

With these two definitions in hand, we can now state the definitions of some of the (statistical, group) fairness measures.

**Predictive Parity**. An algorithm satisfies *predictive parity* (with respect to $T$, $P$) just in case $T \perp\!\!\!\perp P \mid \hat{T}$, *i.e.*, iff

$$\Pr(T \mid P \,\&\, \hat{T}) = \Pr(T \mid \neg P \,\&\, \hat{T}).$$

**True Positive Parity**. An algorithm satisfies *true positive parity* (with respect to $T$, $P$) just in case $\hat{T} \perp\!\!\!\perp P \mid T$, *i.e.*, iff

$$\Pr(\hat{T} \mid P \,\&\, T) = \Pr(\hat{T} \mid \neg P \,\&\, T).$$

**False Positive Parity**. An algorithm satisfies *false positive parity* (with respect to $T$, $P$) just in case $\hat{T} \perp\!\!\!\perp P \mid \neg T$, *i.e.*, iff

$$\Pr(\hat{T} \mid P \,\&\, \neg T) = \Pr(\hat{T} \mid \neg P \,\&\, \neg T).$$

**Statistical Parity**. An algorithm satisfies *statistical parity* (with respect to $T$, $P$) just in case $\hat{T} \perp\!\!\!\perp P$, *i.e.*, iff

$$\Pr(\hat{T} \mid P) = \Pr(\hat{T} \mid \neg P).$$





**COMPAS Data**. In the overall population (all 18293 defendants):

| | $\hat{T}$ (high risk score) | $\neg\hat{T}$ (nonhigh risk score) |
|---|---|---|
| $T$ (actually recidivist) | 2921 ($TP_o$) | 5489 ($FN_o$) |
| $\neg T$ (actually non-recidivist) | 1693 ($FP_o$) | 8190 ($TN_o$) |

True Positive Rate (overall): $\frac{TP_o}{TP_o+FN_o} = 35\%$; False Positive Rate (overall): $\frac{FP_o}{FP_o+TN_o} = 17\%$

In the $P$ sub-population (9779 black defendants), we have:

| | $\hat{T}$ (high risk score) | $\neg\hat{T}$ (nonhigh risk score) |
|---|---|---|
| $T$ (actually recidivist) | 2174 ($TP_P$) | 2902 ($FN_P$) |
| $\neg T$ (actually non-recidivist) | 1226 ($FP_P$) | 3477 ($TN_P$) |

True Positive Rate (black): 43%; False Positive Rate (black): 26%

In the $\neg P$ sub-population (8514 nonblack defendants), we have:

| | $\hat{T}$ (high risk score) | $\neg\hat{T}$ (nonhigh risk score) |
|---|---|---|
| $T$ (actually recidivist) | 747 ($TP_{\neg P}$) | 2587 ($FN_{\neg P}$) |
| $\neg T$ (actually non-recidivist) | 467 ($FP_{\neg P}$) | 4713 ($TN_{\neg P}$) |

True Positive Rate (nonblack): 22%; False Positive Rate (nonblack): 9%

**True Positive Parity** & **False Positive Parity** aren't even *approximately* satisfied here ($0.43 \not\approx 0.22$ and $0.26 \not\approx 0.09$).





More generally, we can use the contingency tables on the previous slide to see whether other fairness measures were (even approximately) satisfied by the COMPAS algorithm.

To test **Predictive Parity**, we need to estimate and compare $\Pr(T \mid P \,\&\, \hat{T})$ and $\Pr(T \mid \neg P \,\&\, \hat{T})$ using the COMPAS data.

$\Pr(T \mid P \,\&\, \hat{T})$ is the proportion of black defendants with high COMPAS scores who actually recidivated, which is:

$$\Pr(T \mid P \,\&\, \hat{T}) = \frac{TP_P}{TP_P + FP_P} = \frac{2174}{2174 + 1226} = \frac{2174}{3400} = 0.64$$

Similarly, $\Pr(T \mid \neg P \,\&\, \hat{T})$ is the proportion of nonblack defendants with high COMPAS scores who actually recidivated:

$$\Pr(T \mid \neg P \,\&\, \hat{T}) = \frac{TP_{\neg P}}{TP_{\neg P} + FP_{\neg P}} = \frac{747}{747 + 467} = \frac{747}{1214} = 0.62$$

As we can see, the COMPAS algorithm *did* (approximately) satisfy **Predictive Parity** (in this population of defendants).





To test **Statistical Parity**, we need to estimate and compare $\Pr(\hat{T} \mid P)$ and $\Pr(\hat{T} \mid \neg P)$ using the COMPAS data.

$\Pr(\hat{T} \mid P)$ is the proportion of black defendants who received high risk scores from COMPAS, which is given by:

$$\Pr(\hat{T} \mid P) = \frac{TP_P + FP_P}{TP_P + FP_P + FN_P + TN_P} = \frac{3400}{9779} = 0.35$$

Similarly, $\Pr(\hat{T} \mid \neg P)$ is the proportion of nonblack defendants who received high risk scores from COMPAS, which is:

$$\Pr(\hat{T} \mid \neg P) = \frac{TP_{\neg P} + FP_{\neg P}}{TP_{\neg P} + FP_{\neg P} + FN_{\neg P} + TN_{\neg P}} = \frac{1214}{8514} = 0.14$$

So, the COMPAS algorithm also failed to (even approximately) satisfy **Statistical Parity**. But, it is worth noting in this connection that **Equal Base Rates** ($T \perp\!\!\!\perp P$) also fails here, since

$$\Pr(T \mid P) = \frac{TP_P + FN_P}{TP_P + FP_P + FN_P + TN_P} = \frac{5076}{9779} = 0.60$$

$$\Pr(T \mid \neg P) = \frac{TP_{\neg P} + FN_{\neg P}}{TP_{\neg P} + FP_{\neg P} + FN_{\neg P} + TN_{\neg P}} = \frac{3334}{8514} = 0.39$$

**Theorem 1** [3]. Unless **Equal Base Rates** ($T \perp\!\!\!\perp P$) holds, it is *impossible* to simultaneously satisfy *all* of: **Predictive Parity**, **True Positive Parity**, **False Positive Parity** & **Statistical Parity**.

More generally [6], as long as $\Pr(T \mid P) \not\approx \Pr(T \mid \neg P)$, no algorithm can — *even approximately* — simultaneously satisfy all four of the fairness measures defined above.

Finally, consider the following pair of weak background assumptions regarding the algorithm being evaluated.

**Imperfection**. An algorithm is *imperfect* iff $\Pr(\hat{T} \mid P \,\&\, \neg T) \neq 0$, $\Pr(\hat{T} \mid \neg P \,\&\, \neg T) \neq 0$, $\Pr(\hat{T} \mid P \,\&\, T) \neq 1$, and $\Pr(\hat{T} \mid \neg P \,\&\, T) \neq 1$.

**Nonzero Precision**. An algorithm has *nonzero precision* iff *either* $\Pr(T \mid \hat{T} \,\&\, P) \neq 0$ *or* $\Pr(T \mid \hat{T} \,\&\, \neg P) \neq 0$.

**Theorem 2** [6]. Assuming **Unequal Base Rates**, no algorithm satisfying both **Imperfection** and **Nonzero Precision** can simultaneously satisfy the three fairness measures: **Predictive Parity**, **True Positive Parity**, and **False Positive Parity**.





Many other fairness measures (that can be defined in terms of the three contingency tables described above) have been proposed, applied, and defended in the literature [7, 2, 8].

We have created a *Mathematica* notebook for exploring impossibility theorems involving fairness measures [4].

Our notebook uses `PrSAT` — a decision procedure for probability calculus which BF implemented in *Mathematica* [5].

There, we study a total of 8 fairness measures (the 4 discussed here, plus 4 more that appear in the recent literature).

We show how to discover new impossibility theorems by automatically checking various subsets of these eight conditions (plus various background assumptions like those above).

In principle, one could use `PrSAT` to discover *all possible* impossibility theorems of these kinds (since it is a general decision procedure for probability calculus). Brief demo...





# Exploring Impossibility Results for Algorithmic Fairness Using PrSAT

Tina Eliassi-Rad & Branden Fitelson
June 5, 2023

## PrSAT — A Decision Procedure for Probability Calculus

**PrSAT** is a decision procedure for probability calculus that has been implemented in *Mathematica* (it's been tested on versions of *Mathematica* up to v12.3). See Fitelson (2008) for details.

The package is self-contained, and can be downloaded from the following website (which also includes instructions for installation and use).

http://fitelson.org/PrSAT/

We begin by loading the **PrSAT** package (which defines all the *Mathematica* functions we'll be using).

```
In[1]:= << PrSAT`
```

## Notation, Fairness Measures, and Side Conditions/Auxiliary Assumptions

We will be discussing binary classification. Our binary classifier $\hat{T}$ can take two values: $\hat{T} = 1$ or $\hat{T} = 0$. We will denote these *predicted* values as $\hat{T}$ and $\neg\hat{T}$, respectively. The *actual* value of the parameter in question will either take the value T = 1 or T = 0, and we will denote these two possibilities as T and ¬T,






[1] J. Angwin *et al.*, *Machine Bias*, *Propublica*, 2016. `http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[2] S. Caton and C. Haas, *Fairness in Machine Learning: A Survey*, 2020.

[3] A. Chouldechova, *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*, 2017.

[4] T. Eliassi-Rad and B. Fitelson, *Exploring Impossibility Theorems for Algorithmic Fairness with* `PrSAT`, 2023. `http://fitelson.org/exploring_impossibility.pdf`.

[5] B. Fitelson, *A Decision procedure for Probability Calculus with Applications*, 2008. `http://fitelson.org/PrSAT/`

[6] J. Kleinberg, S. Mullainathan, and M. Raghavan, *Inherent trade-offs in the fair determination of risk scores*, 2016.

[7] N. Mehrabi *et al.*, *A survey on bias and fairness in machine learning*, 2019.

[8] *Fairness (machine learning)*, *Wikipedia*, Wikimedia Foundation, 2023. `http://en.wikipedia.org/wiki/Fairness_(machine_learning)`.