PHIL 408Q/PHPE 308D
Fairness

Eric Pacuit, University of Maryland

April 11, 2024

Jana Schaich Borg, Walter Sinnott-Armstrong, and Vincent Contizer (2024). *Moral AI: And How We Get There*. Chapter 4: Can AI be fair?, Penguin Books.

Even if AI predictors cannot help but be unfair in some ways, it is still crucial to compare AI predictions to predictions by human judges...

So is AI better than human judges? The discussion comments on the Pro-Publica article framed the issues this way:

So is AI better than human judges? The discussion comments on the Pro-Publica article framed the issues this way:

> **Commenter B**: What is scary is that the results of this program [using COMPAS in Broward County] have been shown to be inaccurate and racially biased (even after controlling for different rates of crimes between certain races).

So is AI better than human judges? The discussion comments on the Pro-Publica article framed the issues this way:

> **Commenter B**: What is scary is that the results of this program [using COMPAS in Broward County] have been shown to be inaccurate and racially biased (even after controlling for different rates of crimes between certain races).
>
> **Commenter K**: Even scarier is when 10,000 judges across the country make decisions where no one can see their 'algorithm' and bias - and we just let them continue to perpetuate injustice. I prefer an algorithm that everyone can see, study, and work to fix. It's easier to fix and test the algorithm than to train and hope judges don't bring bias into decision-making.

At this point there really isn't enough evidence to make definitive conclusions about when human judges or AI systems are more biased, and this comparison might well change with context and as AI develops.

- Even if an AI is less biased, human judges can still be biased in how they apply or reject the AI's recommendations.

Advantages of using AI

- **Potential for transparency**: Many AI systems function as 'black boxes' whose reasons for making predictions are very difficult, if not impossible, to discern. For such reasons, AI predictions are sometimes opaque. Nonetheless, other AI systems are *explainable* and *interpretable*, while still providing good prediction performance.

Advantages of using AI

- **Potential for transparency**: Many AI systems function as 'black boxes' whose reasons for making predictions are very difficult, if not impossible, to discern. For such reasons, AI predictions are sometimes opaque. Nonetheless, other AI systems are *explainable* and *interpretable*, while still providing good prediction performance.
- **Explicit prejudice and indirect proxies**: AIs can be intentionally designed to avoid using racial or other demographic categories in its predictions...

Advantages of using AI

- **Potential for transparency**: Many AI systems function as 'black boxes' whose reasons for making predictions are very difficult, if not impossible, to discern. For such reasons, AI predictions are sometimes opaque. Nonetheless, other AI systems are *explainable* and *interpretable*, while still providing good prediction performance.
- **Explicit prejudice and indirect proxies**: AIs can be intentionally designed to avoid using racial or other demographic categories in its predictions... Unfortunately, even if an AI is not given racial information directly, the data that it analyses can still include information about other categories that are highly correlated with racial categories (called 'proxies').

Advantages of using AI

- **Potential for transparency**: Many AI systems function as 'black boxes' whose reasons for making predictions are very difficult, if not impossible, to discern. For such reasons, AI predictions are sometimes opaque. Nonetheless, other AI systems are *explainable* and *interpretable*, while still providing good prediction performance.
- **Explicit prejudice and indirect proxies**: AIs can be intentionally designed to avoid using racial or other demographic categories in its predictions... Unfortunately, even if an AI is not given racial information directly, the data that it analyses can still include information about other categories that are highly correlated with racial categories (called 'proxies').
- **Corrections and protected classes**: In theory, AI algorithms should be able to leverage their quantitative models of the world to statistically correct for certain unfair outcomes, at least to some degree.

Procedural justice

Even if AI optimists win out and the legal system ends up using AIs that are shown to be sufficiently fair *distributively* and *retributively*, could those same AIs still be *procedurally* unjust or unfair?

Among other things (such as an impartial judge and a speedy trial), procedural justice in law is usually thought to require a right for each side to cross-examine the other's witnesses and, more generally, to question their evidence.

Among other things (such as an impartial judge and a speedy trial), procedural justice in law is usually thought to require a right for each side to cross-examine the other's witnesses and, more generally, to question their evidence.

Each side must be able to understand the other's witnesses and evidence for any cross-examination to be effective.

This ability to question becomes a critical issue when AI predictions are a basis for legal decisions.

Among other things (such as an impartial judge and a speedy trial), procedural justice in law is usually thought to require a right for each side to cross-examine the other's witnesses and, more generally, to question their evidence.

Each side must be able to understand the other's witnesses and evidence for any cross-examination to be effective.

This ability to question becomes a critical issue when AI predictions are a basis for legal decisions.

If the AIs that made those predictions are unintelligible to anyone other than an AI expert, or if they are impossible even for experts to understand, then the defense loses its ability to respond effectively. That would make court procedures unfair.

Loomis v. Wisconsin

- Eric Loomis was charged with taking part in a drive-by shooting. He denied firing the shots but pleaded guilty to 'attempting to flee a traffic officer and operating a motor vehicle without the owner's consent'.

Loomis v. Wisconsin

- Eric Loomis was charged with taking part in a drive-by shooting. He denied firing the shots but pleaded guilty to 'attempting to flee a traffic officer and operating a motor vehicle without the owner's consent'.
- Before a COMPAS score was introduced into Loomis's case, the prosecution and defense had agreed upon a plea deal of one year in county jail with probation.

Loomis v. Wisconsin

- Eric Loomis was charged with taking part in a drive-by shooting. He denied firing the shots but pleaded guilty to 'attempting to flee a traffic officer and operating a motor vehicle without the owner's consent'.
- Before a COMPAS score was introduced into Loomis's case, the prosecution and defense had agreed upon a plea deal of one year in county jail with probation.
- At sentencing, a probation officer shared that the COMPAS AI predicted Looms would probably reoffend.

Loomis v. Wisconsin

- The trial judge stated, You're identified, through the COMPAS assessment, as an individual who is at high risk to the community.

  In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend.

Loomis v. Wisconsin

- The trial judge stated, You're identified, through the COMPAS assessment, as an individual who is at high risk to the community.

  In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend.

- Loomis was then sentenced to six years in prison and five years of extended supervision.

Loomis v. Wisconsin

Loomis appealed the sentencing decision.

Loomis v. Wisconsin

Loomis appealed the sentencing decision.

One of his critical arguments was that his trial was unfair not only because COMPAS was unfair to certain groups, but also because COMPAS's predictive model was both proprietary and complicated (being based on 137 questions), so there was no realistic way for Loomis or his attorney to know how or why COMPAS arrived at its risk prediction or to cross-examine, understand, or respond to its prediction.

Loomis v. Wisconsin

Loomis appealed the sentencing decision.

One of his critical arguments was that his trial was unfair not only because COMPAS was unfair to certain groups, but also because COMPAS's predictive model was both proprietary and complicated (being based on 137 questions), so there was no realistic way for Loomis or his attorney to know how or why COMPAS arrived at its risk prediction or to cross-examine, understand, or respond to its prediction.

Loomis ultimately lost his appeal, but many legal scholars think he should have won, particularly because of this procedural argument.

The procedural right to know why and how COMPAS is labelling people as ‘likely to reoffend’ is important not only to defendants.

The procedural right to know why and how COMPAS is labelling people as 'likely to reoffend' is important not only to defendants.

COMPAS's inner workings are important for judges as well.

- The trial judge in Loomis's case needed to be able to make informed decisions about when (and how much) to trust COMPAS's predictions in order to be justified in believing that Looms was truly 'extremely high risk to re-offend'.
- Without this knowledge, the judge would need to accept or reject the algorithm's prediction blindly and could end up confidently following the prediction even when it is unreliable.

Does interpretability solve the problem?

Algorithms are considered interpretable when humans can figure out what caused them to produce their outputs.

Does interpretability solve the problem?

Algorithms are considered interpretable when humans can figure out what caused them to produce their outputs.

If we required all AIs used in the justice system to be interpretable, and also required the developers of such AIs to share how their AIs were trained and how they work, would that remove all concerns about the procedural justice of these AIs?

Does interpretability solve the problem?

- Black-box deep learning AIs are popular because they often perform better than any other currently known AI technique. Interpretable algorithms are sometimes less accurate than uninterpretable algorithms, and this inaccuracy really matters when it comes to decisions that can affect whether somebody will be put in jail and for how long.

Does interpretability solve the problem?

- Black-box deep learning AIs are popular because they often perform better than any other currently known AI technique. Interpretable algorithms are sometimes less accurate than uninterpretable algorithms, and this inaccuracy really matters when it comes to decisions that can affect whether somebody will be put in jail and for how long.

- Another complication is that 'interpretability' means different things to different people.
  - Even if a computer scientist can understand and predict how an interpretable algorithm will behave, that doesn't mean a typical lawyer or defendant will be able to understand it or predict its behavhour.
  - What kind and what degree of intelligibility is required for a fair legal system?

The silver lining in all of this is that the introduction of AI across so many aspects of life has helped to make more of us aware of many forms of injustice in the decisions that humans have traditionally made.

The silver lining in all of this is that the introduction of AI across so many aspects of life has helped to make more of us aware of many forms of injustice in the decisions that humans have traditionally made.

Even if we haven't yet figured out how to apply AI fairly in all circumstances, at least AI is highlighting unfairness that needs to be addressed.

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

Sina Fazelpour and David Danks (2021). *Algorithmic bias: Senses, sources, solutions*. 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.

Example: Hiring

- Suppose that applicants for a job are from two different groups, “male” and “female.”

Example: Hiring

- Suppose that applicants for a job are from two different groups, "male" and "female."
- Every applicant is either qualified or not, but this is not directly observable. Rather, each applicant has taken a test, and the result of this test for applicant is positively correlated with whether he or she is qualified. To make things concrete, suppose that the test is scored on a $0 - 100$ point scale.

Example: Hiring

- Suppose that applicants for a job are from two different groups, "male" and "female."
- Every applicant is either qualified or not, but this is not directly observable. Rather, each applicant has taken a test, and the result of this test for applicant is positively correlated with whether he or she is qualified. To make things concrete, suppose that the test is scored on a $0 - 100$ point scale.
- The employer can observe both the applicant's test score and his or her group membership, and suppose that the employer hires any applicant from group $g \in \{male, female\}$ if and only if his or her test score is greater than or equal to the employer's threshold for group g, denoted by $t(g) \in \{0, \ldots, 100, 101\}$

Example: Hiring

Both AF and SD are interested in the pair of thresholds used by the employer, $t(male)$ and $t(female)$.

This stylized setting allows us to clearly identify discrimination between the two groups: whenever $t(male) \neq t(female)$

Example: Hiring

Studies of Algorithmic Fairness tend to focus on questions like:

1. How do the thresholds affect the applicants' welfares?
2. What does it mean to treat applicants from both groups of applicants fairly?
3. Which pair(s) of thresholds (if any) treat both groups of applicants fairly?

Example: Hiring

Studies of Algorithmic Fairness tend to focus on questions like:

1. How do the thresholds affect the applicants' welfares?
2. What does it mean to treat applicants from both groups of applicants fairly?
3. Which pair(s) of thresholds (if any) treat both groups of applicants fairly?

Studies of Statistical Discrimination tend to focus on questions like:

1. How do the thresholds affect the employer's welfare?
2. Which pair(s) of thresholds maximize the employer's welfare?
3. What factors might justify the employer using different thresholds for the two groups?
4. How do these thresholds affect individual and group behavior?

Traffic Cameras, Fairness, and Discrimination

≡ PROPUBLICA Donate

Chicago's "Race-Neutral" Traffic Cameras Ticket Black and Latino Drivers the Most

A ProPublica analysis found that traffic cameras in Chicago disproportionately ticket Black and Latino motorists. But city officials plan to stick with them — and other cities may adopt them too.

by Emily Hopkins and Melissa Sanchez

Jan. 11, 2022, 5 a.m. EST

Traffic Cameras, Fairness, and Discrimination

The study found that, in Chicago in 2020, "the ticketing rate for households in majority-Black ZIP codes jumped to more than three times that of households in majority-white areas. For households in majority-Hispanic ZIP codes, there was an increase, but it was much smaller."

Traffic Cameras, Fairness, and Discrimination

An Algorithmic Fairness perspective on this situation essentially asks why this disparity emerges and, more provocatively, how one might reduce or eliminate it.

Traffic Cameras, Fairness, and Discrimination

An Algorithmic Fairness perspective on this situation essentially asks why this disparity emerges and, more provocatively, how one might reduce or eliminate it.

This perspective is particularly helpful in this type of setting because, while this disparity has widened over the two decades since the cameras were introduced in Chicago, there is little reason to suspect that traffic cameras themselves are distinguishing between drivers based on their race or home neighborhood, per se.

Traffic Cameras, Fairness, and Discrimination

An Algorithmic Fairness perspective on this situation essentially asks why this disparity emerges and, more provocatively, how one might reduce or eliminate it.

This perspective is particularly helpful in this type of setting because, while this disparity has widened over the two decades since the cameras were introduced in Chicago, there is little reason to suspect that traffic cameras themselves are distinguishing between drivers based on their race or home neighborhood, per se.

In this specific case, this perspective allows one to see that the disparity is at least arguably due to speed limits and driving conditions being distributed in a "non-race blind" fashion across Chicago.

Traffic Cameras, Fairness, and Discrimination

Chicago Mayor Lori Lightfoot's administration described traffic cameras as "a tool in the toolkit to help alleviate" traffic fatalities and, from an empirical standpoint, Black Chicagoans were twice as likely to die in a traffic accident as white Chicagoans in 2017. Accordingly, Black Chicagoans are differentially treated by both traffic accidents and traffic tickets.

Traffic Cameras, Fairness, and Discrimination

From a Statistical Discrimination standpoint, on the other hand, one might ask why Chicago is using traffic cameras, in spite of the clear racial disparity in which citizens receive tickets.

Traffic Cameras, Fairness, and Discrimination

From a Statistical Discrimination standpoint, on the other hand, one might ask why Chicago is using traffic cameras, in spite of the clear racial disparity in which citizens receive tickets.

As the ProPublica article describes, Chicago Mayor Lori Lightfoot proposed lowering the minimum speed at which a speeding ticket would be issued.

Traffic Cameras, Fairness, and Discrimination

From a Statistical Discrimination standpoint, on the other hand, one might ask why Chicago is using traffic cameras, in spite of the clear racial disparity in which citizens receive tickets.

As the ProPublica article describes, Chicago Mayor Lori Lightfoot proposed lowering the minimum speed at which a speeding ticket would be issued.

This proposal, which was adopted by the Chicago City Council in 2021, prompted some to question how much Mayor Lightfoot cared about racial disparities, as opposed to the City of Chicago's serious structural deficit.

Traffic Cameras, Fairness, and Discrimination

From a Statistical Discrimination standpoint, on the other hand, one might ask why Chicago is using traffic cameras, in spite of the clear racial disparity in which citizens receive tickets.

As the ProPublica article describes, Chicago Mayor Lori Lightfoot proposed lowering the minimum speed at which a speeding ticket would be issued.

This proposal, which was adopted by the Chicago City Council in 2021, prompted some to question how much Mayor Lightfoot cared about racial disparities, as opposed to the City of Chicago's serious structural deficit.

The question of "is Mayor Lightfoot more interested in racial equality or city revenue?" is directly analogous to the seminal question in statistical discrimination, "is that employer simply maximizing profits or are they racist?"

Algorithmic Fairness vs. Statistical Discrimination

In terms of the traffic camera example, we can also distinguish the AF and SD viewpoints as

- Algorithmic Fairness: Can we make Chicago's traffic enforcement more fair? If so, how?
- Statistical Discrimination: Why did Chicago use an unfair traffic enforcement algorithm?

Rationality vs. Fairness

A theme running throughout this article is that a key contrast between the AF & SD approaches revolves around the question of rationality or, in slightly different terms, *efficiency*.

Many SD theories are focused on how the pursuit of efficiency (e.g., by an employer, job applicant, government, or other individuals) can generate behavior that is discriminatory.

On the other hand, AF is less concerned with efficiency (partly because the basic framework does not presume anything about individuals' motives/goals).

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** $x_i = (x_1^1, \ldots, x_i^m)$

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

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

1. A profile of **permissible traits** $x_i = (x_1^1, \ldots, x_i^m)$

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

2. A profile of **sensitive traits** $a_i, ai = (x_1^1, \ldots, x_i^k)$.

   Examples: Gender, race, ethnicity, marital status.

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

1. A profile of **permissible traits** $x_i = (x_1^1, \ldots, x_i^m)$

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

2. A profile of **sensitive traits** $a_i, ai = (x_1^1, \ldots, x_i^k)$.

   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** $x_i = (x_1^1, \ldots, x_i^m)$

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

2. A profile of **sensitive traits** $a_i, ai = (x_1^1, \ldots, x_i^k)$.

   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 ($\delta_i = 1$)	Positive ($\delta_i = 0$)	
Positive ($y_i = 1$)	True Positive (TP)	False Negative (FN)	TPR: $\frac{TP}{TP+FN}$
Negative ($y_i = 0$)	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

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

$$0.348 \not\approx 0.143$$

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.

Impossibility Theorem

Theorem. If an algorithm satisfies Predictive Parity and Error Rate Balance, then one or both of the following must be satisfied:

- Perfect Predictor: $Pr[y_i = 1 \mid a_i, x_i] \in \{0, 1\}$ for all x_i, a_i
- Equal Base Rates: $Pr[y_i = 1 \mid a_i] = Pr[y_i \mid a_i']$ for all i, a_i'

J. Kleinberg, S. Mullainathan, and M. Raghavan (2016). *Inherent trade-offs in the fair determination of risk scores*. https://arxiv.org/abs/1609.05807.

A. Chouldechova (2017). *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*. https://arxiv.org/abs/1610.07524.

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.