PHPE 308M/PHIL 209F
Fairness

Eric Pacuit, University of Maryland

November 12, 2025

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 Loomis 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.

COMPAS: Evaluating Risk of Recidivism

https://embed.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE/

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.

Surveys

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.

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 Criterion

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

Fairness Criterion

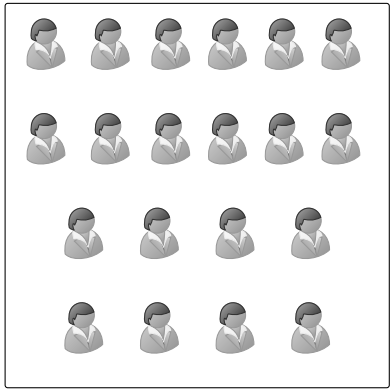
Some fairness criterion involve studying the internal workings of the algorithm. E.g., the algorithm cannot be based on certain features.

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

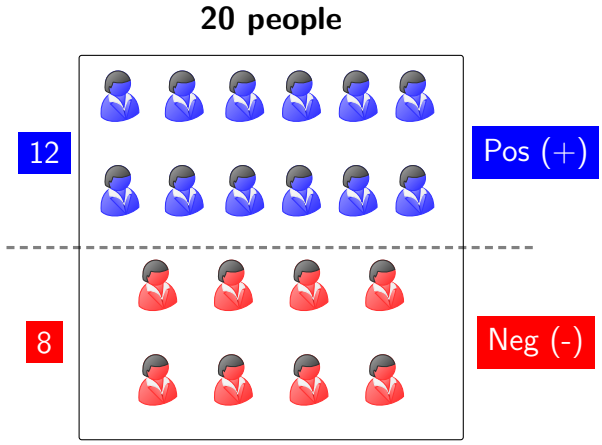
The criteria can be evaluated without actually looking at the inner workings of the algorithm, which may be proprietary or otherwise opaque. Instead, we just have look at the results—what the algorithm predicted and what actually happened.

Example

20 people

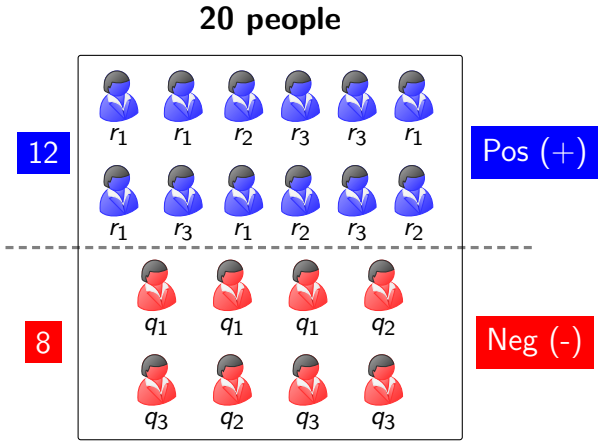


Example



Binary predictions: 12 classified as positive (Pos); 8 classified as negative (Neg)

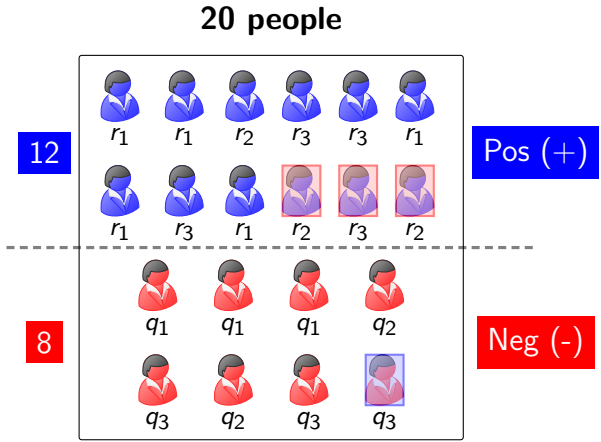
Example



Binary predictions: 12 classified as positive (Pos); 8 classified as negative (Neg)

Predict Risk Scores: $0 \leq q_1, q_2, q_3, r_1, r_2, r_3 \leq 1$

Example

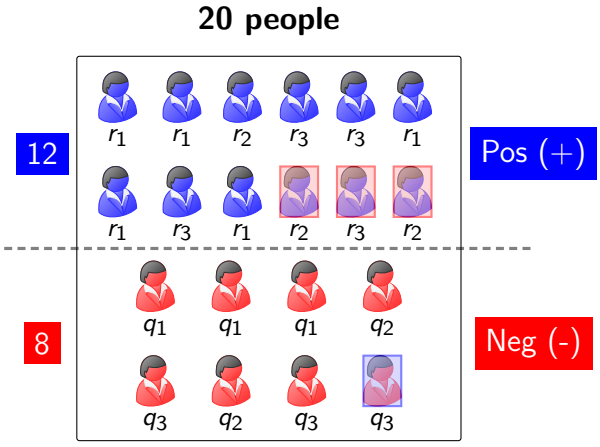


Binary predictions: 12 classified as positive (Pos); 8 classified as negative (Neg)

Predict Risk Scores: $0 \leq q_1, q_2, q_3, r_1, r_2, r_3 \leq 1$

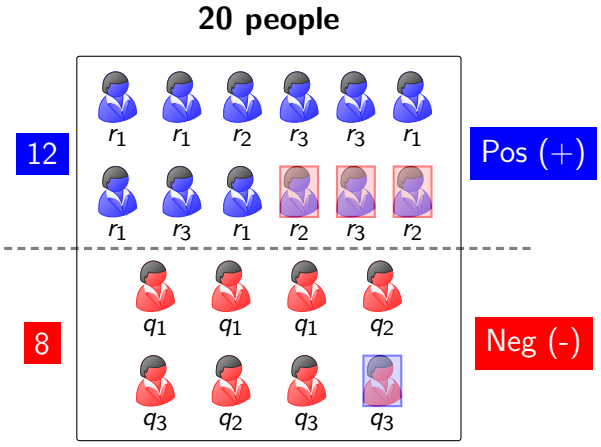
Actuality: 3 classified as Pos are misclassified, 1 classified as Neg is misclassified

Confusion Matrix



	Pred. +	Pred. −
Actual +	9 TP	1 FN
Actual −	3 FP	7 TN

Confusion Matrix

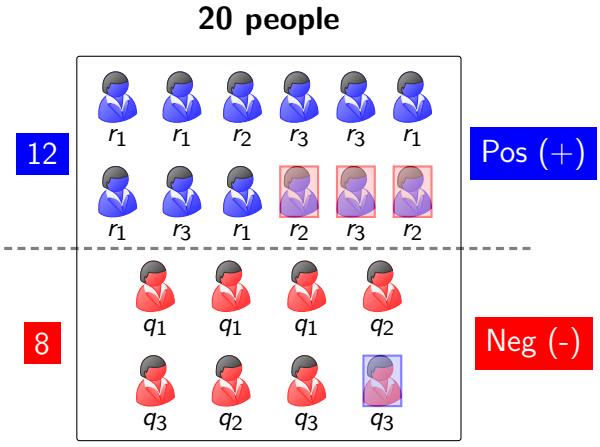


	Pred. +	Pred. −
Actual +	9 TP	1 FN
Actual −	3 FP	7 TN

Accuracy: $\frac{9+7}{20} = \frac{4}{5}$

Base Rate: $\frac{9+1}{20} = \frac{1}{2}$

Confusion Matrix



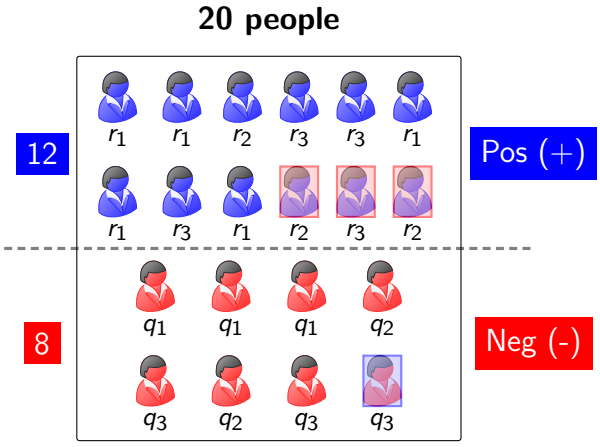
	Pred. +	Pred. −
Actual +	9 TP	1 FN
Actual −	3 FP	7 TN

Error Rates:

False Neg. Rate: $\frac{1}{1+9} = \frac{1}{10}$

False Pos. Rate: $\frac{3}{3+7} = \frac{3}{10}$

Confusion Matrix



	Pred. +	Pred. −
Actual +	9 TP	1 FN
Actual −	3 FP	7 TN

Predictive Value:

Pos. Predictive Value: $\frac{9}{9+3} = \frac{3}{4}$

Neg. Predictive Value: $\frac{7}{1+7} = \frac{7}{8}$

Confusion Matrix

	Pred. +	Pred. −
Actual +	9 TP	1 FN
Actual −	3 FP	7 TN

Error Rates:

False Neg. Rate: $\frac{1}{1+9} = \frac{1}{10}$

False Pos. Rate: $\frac{3}{3+7} = \frac{3}{10}$

Given the truth, how often is the prediction wrong?

Predictive Value:

Pos. Predictive Value: $\frac{9}{9+3} = \frac{3}{4}$

Neg. Predictive Value: $\frac{7}{1+7} = \frac{7}{8}$

Given the prediction, how often is the prediction correct?

Compas Data

Overall population (18,293 defendants)

	Pred: High Risk	Pred: Not High Risk
Actual Recidivist	2921 TP	5489 FN
Actual Non-Recid.	1693 FP	8190 TN

Accuracy: $\frac{2921+8190}{2921+5489+1693+8190} \approx 0.607$

Base Rate: $\frac{2921+5489}{2921+5489+1693+8190} \approx 0.459$

Black Defendants (n = 9,779)

	Pred: High Risk	Pred: Not High Risk
Actual Recidivist	2174 TP	2902 FN
Actual Non-Recid.	1226 FP	3477 TN

Non-Black Defendants (n = 8,514)

	Pred: High Risk	Pred: Not High Risk
Actual Recidivist	747 TP	2587 FN
Actual Non-Recid.	467 FP	4713 TN

Predictive Parity

Black Defendants

	Pred +	Pred –
Actual +	2174	2902
Actual –	1226	3477

Non-Black Defendants

	Pred +	Pred –
Actual +	747	2587
Actual –	467	4713

Predictive Parity

Black Defendants

	Pred +	Pred –
Actual +	2174	2902
Actual –	1226	3477

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

Non-Black Defendants

	Pred +	Pred –
Actual +	747	2587
Actual –	467	4713

Predictive Parity

Black Defendants

	Pred +	Pred –
Actual +	2174	2902
Actual –	1226	3477

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

Non-Black Defendants

	Pred +	Pred –
Actual +	747	2587
Actual –	467	4713

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

Predictive Parity

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

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

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

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

$$0.639 \approx 0.615$$

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

Error Rate Balance

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

Error Rate Balance

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

$$FNR = \frac{2902}{2174+2902} \approx 0.572$$

$$FPR = \frac{1226}{1226+3477} \approx 0.261$$

Error Rate Balance

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

$$FNR = \frac{2902}{2174+2902} \approx 0.572$$
$$FPR = \frac{1226}{1226+3477} \approx 0.261$$

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

$$FNR = \frac{2587}{747+2587} \approx 0.776$$
$$FPR = \frac{467}{467+4713} \approx 0.090$$

Error Rate Balance

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

$$FNR = \frac{2902}{2174+2902} \approx 0.572$$
$$FPR = \frac{1226}{1226+3477} \approx 0.261$$

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

$$FNR = \frac{2587}{747+2587} \approx 0.776$$
$$FPR = \frac{467}{467+4713} \approx 0.090$$

$$0.572 \not\approx 0.776 \text{ and } 0.261 \not\approx 0.090$$

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

Base Rates

Black Defendants

	Pred +	Pred –
Actual +	2174	2902
Actual –	1226	3477

Non-Black Defendants

	Pred +	Pred –
Actual +	747	2587
Actual –	467	4713

Base Rates

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

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

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

Base Rates

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

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

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

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

Base Rates

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

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

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

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

$$0.519 \not\approx 0.392$$

The base rates of recidivism are not equal.

Demographic Parity (Statistical Parity)

Black Defendants

	Pred +	Pred –
Actual +	2174	2902
Actual –	1226	3477

Non-Black Defendants

	Pred +	Pred –
Actual +	747	2587
Actual –	467	4713

Demographic Parity (Statistical Parity)

Black Defendants

	Pred +	Pred –
Actual +	2174	2902
Actual –	1226	3477

Non-Black Defendants

	Pred +	Pred –
Actual +	747	2587
Actual –	467	4713

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

Demographic Parity (Statistical Parity)

Black Defendants

	Pred +	Pred –
Actual +	2174	2902
Actual –	1226	3477

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

Non-Black Defendants

	Pred +	Pred –
Actual +	747	2587
Actual –	467	4713

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

Demographic Parity (Statistical Parity)

Black Defendants

	Pred +	Pred –
Actual +	2174	2902
Actual –	1226	3477

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

Non-Black Defendants

	Pred +	Pred –
Actual +	747	2587
Actual –	467	4713

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

$$0.348 \not\approx 0.143$$

Demographic parity requires that the algorithm predicts high risk at equal rates across groups, regardless of sensitive attributes. Here, Black defendants are flagged as high risk 2.4 times more often than non-Black defendants.

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