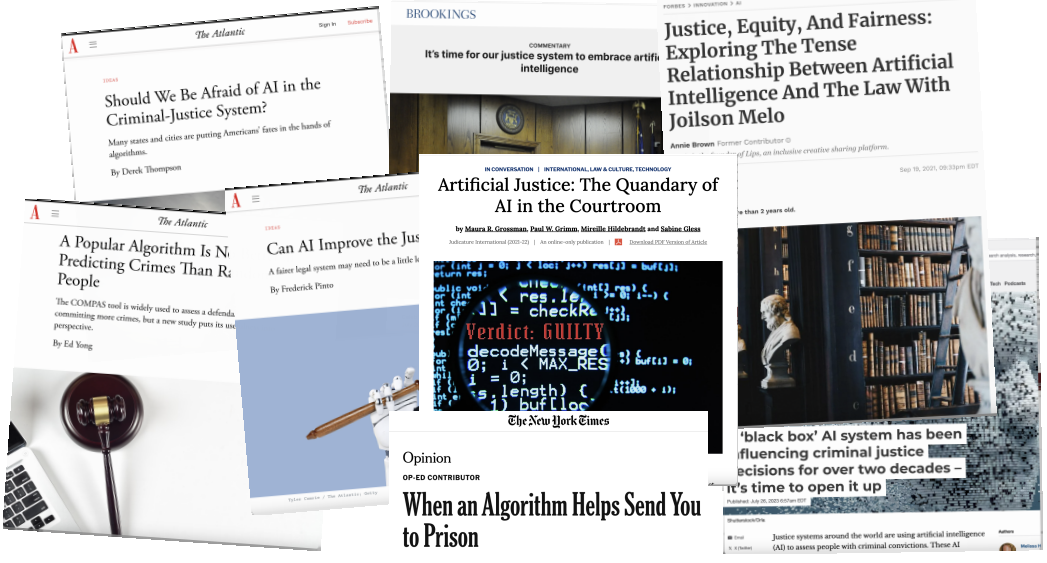
PHIL 408Q/PHPE 308D
Fairness

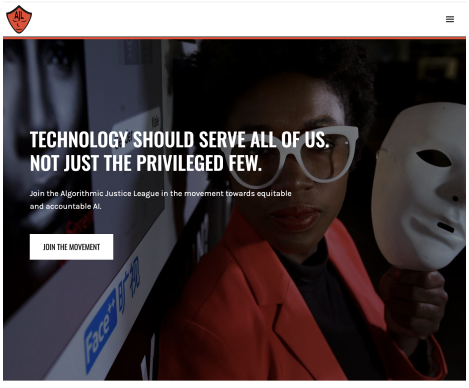
Eric Pacuit, University of Maryland

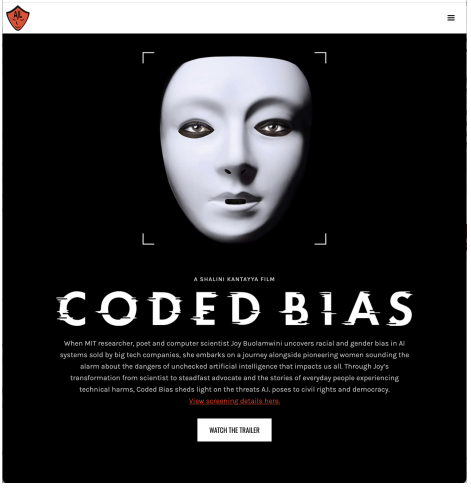
April 9, 2024

Fairness in AI

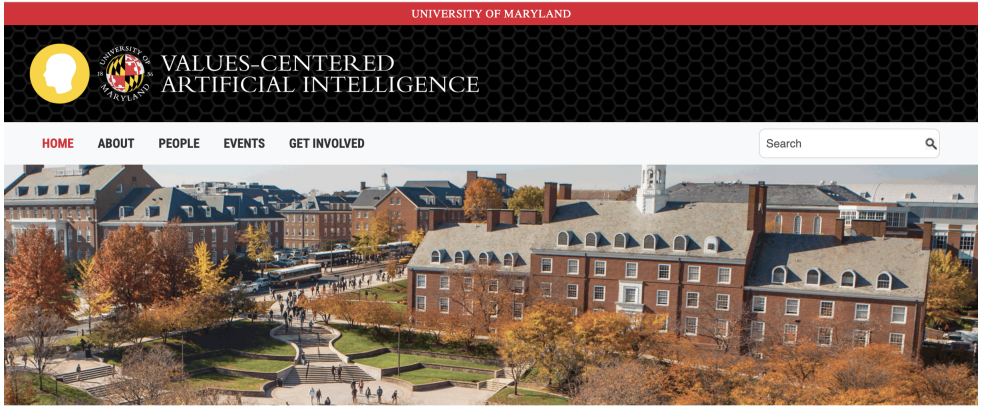

The Atlantic
Sign In
Subscribe
IDEAS
Should We Be Afraid of AI in the Criminal-Justice System?
Many states and cities are putting Americans' fates in the hands of algorithms.
By Derek Thompson
BROOKINGS
COMMENTARY
It's time for our justice system to embrace artifi
intelligence
FORBES > INNOVATION > AI
Justice, Equity, And Fairness: Exploring The Tense Relationship Between Artificial Intelligence And The Law With Joilson Melo
Annie Brown Former Contributor
Sep 19, 2021, 09:33pm EDT
The Atlantic
A Popular Algorithm Is N
Predicting Crimes Than R
People
The COMPAS tool is widely used to assess a defenda
committing more crimes, but a new study puts its us
perspective.
By Ed Yong
The Atlantic
IDEAS
Can AI Improve the Jus
A fairer legal system may need to be a little l
By Frederick Pinto
IN CONVERSATION | INTERNATIONAL, LAW & CULTURE, TECHNOLOGY
Artificial Justice: The Quandary of AI in the Courtroom
by Maura R. Grossman, Paul W. Grimm, Mireille Hildebrandt and Sabine Gless
Judicature International (2021-22) | An online-only publication | Download PDF Version of Article
Verdict: GUILTY
decodeMessage
The New York Times
Opinion
OP-ED CONTRIBUTOR
When an Algorithm Helps Send You to Prison
By Ellora Thadaney Israni
Oct. 26, 2017
than 2 years old.
'black box' AI system has been
fluencing criminal justice
ecisions for over two decades –
it's time to open it up
Published July 26, 2023 6:57am EDT
Justice systems around the world are using artificial intelligence (AI) to assess people with criminal convictions. These AI
Authors

Algorithmic Justice League





AI and Fairness @ UMD



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.

Headlines frequently suggest that AI is unfair to disadvantaged groups in various ways. AI commonly used for hiring, firing, promotion, home loans, and business loans often disfavour Black, female, immigrant, poor, disabled, and neurodiverse applicants, among other groups.

Headlines frequently suggest that AI is unfair to disadvantaged groups in various ways. AI commonly used for hiring, firing, promotion, home loans, and business loans often disfavour Black, female, immigrant, poor, disabled, and neurodiverse applicants, among other groups.

...[G]ood or bad consequences are awarded disproportionately to certain groups of people, usually in the form of harms to already-disadvantaged groups and benefits to already privileged groups. When such biases are unjustified, as they usually are, they are considered to be unfair or unjust - terms that we will use interchangeably.

But if AI is so ‘intelligent’, shouldn’t it know better than to be biased?

But if AI is so ‘intelligent’, shouldn’t it know better than to be biased?

For all the many surprising advances that AI technology makes,
this is one of the arenas where it continues to struggle.

“bias in, bias out”

1. It is very difficult (and often expensive) to assemble data sets that have all demographic groups and interests represented equally, and trained models are usually more accurate at making predictions about groups that are well represented in its training data than groups that are not.

“bias in, bias out”

1. It is very difficult (and often expensive) to assemble data sets that have all demographic groups and interests represented equally, and trained models are usually more accurate at making predictions about groups that are well represented in its training data than groups that are not.

2. A more general reason AIs end up biased is that humans and human social structures are often biased, and our biases are readily built into the AIs we design and create.

"bias in, bias out"

1. It is very difficult (and often expensive) to assemble data sets that have all demographic groups and interests represented equally, and trained models are usually more accurate at making predictions about groups that are well represented in its training data than groups that are not.

2. A more general reason AIs end up biased is that humans and human social structures are often biased, and our biases are readily built into the AIs we design and create. Every time a human decides what data to collect, labels a data point, decides what information should be fed into an AI algorithm, chooses a goal for an AI to pursue, decides how to evaluate an AI model's performance, or decides how to respond to an AI prediction, opportunities are created for our own human biases to be reflected in an AI.

These two overarching causes for AI bias are so pervasive and challenging that most experts, regardless of their level of technologic optimism, agree that AI systems (like humans) are almost never perfectly just or fair.

These two overarching causes for AI bias are so pervasive and challenging that most experts, regardless of their level of technologic optimism, agree that AI systems (like humans) are almost never perfectly just or fair.

This raises the critical questions:

- Should we use AI when we know that it can contribute to injustice?
- Is there perhaps some hope of designing AI systems that would actually reduce injustice, perhaps even in settings where AI currently does not play any role?

Distributive justice

Distributive justice concerns how burdens and benefits are distributed among individuals and groups.

Distributive justice

Distributive justice concerns how burdens and benefits are distributed among individuals and groups.

It seems unfair or unjust for businesses to refuse to hire applicants from a disfavoured group, for municipalities to provide better schools or more police protection to a favoured group, or for countries to require or allow only some groups and not others to serve in the military.

Such practices might be reasonable in certain circumstances, but justifying such inequality would take at least some special reason.

Retributive justice

Retributive justice, in contrast, concerns whether a punishment fits the crime, or, more generally, whether people get what they deserve.

Retributive justice

Retributive justice, in contrast, concerns whether a punishment fits the crime, or, more generally, whether people get what they deserve.

Punishments can be unfair by being too harsh or too lenient. It seems unfair to sentence a car thief to life in prison, because that punishment is too harsh for that crime. On the other hand, it also seems unfair to sentence a rapist to only one day in jail, because that minor punishment is too lenient for such a horrible offence.

Procedural justice

Procedural justice concerns whether the processes or procedures used to reach decisions about how to distribute benefits and burdens are fair.

Procedural justice

Procedural justice concerns whether the processes or procedures used to reach decisions about how to distribute benefits and burdens are fair.

Even a murderer who confesses and is clearly guilty still deserves a fair trial. Similarly, a procedure for selecting political leaders would be unfair if certain races or genders were denied the right to vote, even if the same candidates would win anyway.

The police make over 7 million arrests every year in the US. After arrest and booking comes an arraignment, where a criminal defendant appears in court to hear the charges against them and submit a plea. This arraignment is typically combined with a bail hearing, in which a judge decides where the defendant will live while waiting for the next hearing or trial.

Bail

- The judge can decide to let the defendant go home (or wherever they want) with only a written promise that they will return at the next required court date.
- The judge can also require the defendant to stay in jail during that time if they think the defendant is likely to fail to show for their court appointment or commit a crime in the meantime.
- An intermediate option is to allow the defendant to go home until their next required court appearance if, and only if, they pay a certain amount of money as a security deposit to help ensure they will return for their scheduled court dates.

We will refer to the decision of where a defendant should reside under which conditions while waiting for trial as a 'bail decision'.

Importantly, judges in the United States are not supposed to make these bail decisions on the basis of whether they think the defendant is guilty. Assessments of guilt come later, during the trial.

Instead, judges are typically supposed to base their bail decisions solely on two predictions of what the defendant will do if released: will this defendant flee and fail to appear at the trial? Will this defendant commit another crime while out on bail?

The time pressure makes it unrealistic for judges to ponder or even familiarize themselves with all the relevant details of each case. The time pressure may also make it more likely that judges will rely on some of their documented implicit bias towards or against certain groups when making decisions.

The time pressure makes it unrealistic for judges to ponder or even familiarize themselves with all the relevant details of each case. The time pressure may also make it more likely that judges will rely on some of their documented implicit bias towards or against certain groups when making decisions.

Thus courtrooms across the United States have turned to AI for assistance because they believe that AI can make more accurate predications from complex information and show less bias than humans.

Human judges vs. AI

Responsible actors in every sentences system - from prosecutors to judges to parole officials - make daily judgements about....the risks of recidivism posed by offenders. These judgement, pervasive as they are are notoriously imperfect. They often derive from the intuitions and abilities of individual decisionmakers, who typically lack professional training in the sciences of human behavior...

Human judges vs. AI

Responsible actors in every sentences system - from prosecutors to judges to parole officials - make daily judgements about....the risks of recidivism posed by offenders. These judgement, pervasive as they are are notoriously imperfect. They often derive from the intuitions and abilities of individual decisionmakers, who typically lack professional training in the sciences of human behavior... Actuarial - or statistical - predictions of risk, derived from objective criteria, have been found superior to clinical predictions built on the professional training, experience, and judgment of the persons making predictions.

American Law Institute. *Model Penal Code Sentencing*. 2017: article 6B.09, comment a, 387-389.

In one study looking at bail decisions in New York City, the defendants whom an AI classified as risky failed to appear for trial 56 per cent of the time, committed other new crimes 63 percent of the time, and even committed the most serious crimes (murder, rape, and robbery) 5 percent of the time - all much more than defendants whom the AI did not classify as risky.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018). *Human Decisions and Machine Predictions*. The Quarterly Journal of Economics, 133(1), pp. 237 - 293.

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend

In forecasting who would reoffend, the algorithm made mistakes with Black and White defendants at roughly the same rate but in very different ways.

In forecasting who would reoffend, the algorithm made mistakes with Black and White defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag Black defendants as future criminals, wrongly labeling them this way at almost twice the rate as White defendants.

In forecasting who would reoffend, the algorithm made mistakes with Black and White defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag Black defendants as future criminals, wrongly labeling them this way at almost twice the rate as White defendants.
- White defendants were mislabeled as low risk more often than Black defendants.

In forecasting who would reoffend, the algorithm made mistakes with Black and White defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag Black defendants as future criminals, wrongly labeling them this way at almost twice the rate as White defendants.
- White defendants were mislabeled as low risk more often than Black defendants.

Could this disparity be explained by defendants' prior crimes or the type of crimes they were arrested for?

In forecasting who would reoffend, the algorithm made mistakes with Black and White defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag Black defendants as future criminals, wrongly labeling them this way at almost twice the rate as White defendants.
- White defendants were mislabeled as low risk more often than Black defendants.

Could this disparity be explained by defendants' prior crimes or the type of crimes they were arrested for? No. We ran a statistical test that isolated the effect of race from criminal history and recidivism, as well as from defendants' age and gender. Black defendants were still 77 per cent more likely to be pegged as at higher risk of committing a future violent crime and 45 per cent more likely to be predicted to commit a future crime of any kind.

The first bullet point says that COMPAS has a higher rate of false positives (the percentage predicted to recidivate who did not actually recidivate) for Black defendants than for White.

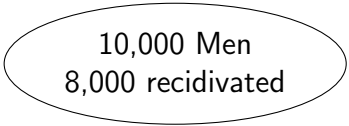
The second bullet point then reports that COMPAS has a higher rate of false negatives (the percentage predicted not to recidivate who did actually recidivate) for White defendants than for Black.

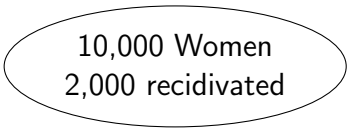
Northpointe, the producer of COMPAS, admitted this difference in mistake rates. However, they replied by showing that COMPAS predictions are still equally accurate on average for Black and for White defendants.

Northpointe, the producer of COMPAS, admitted this difference in mistake rates. However, they replied by showing that COMPAS predictions are still equally accurate on average for Black and for White defendants.

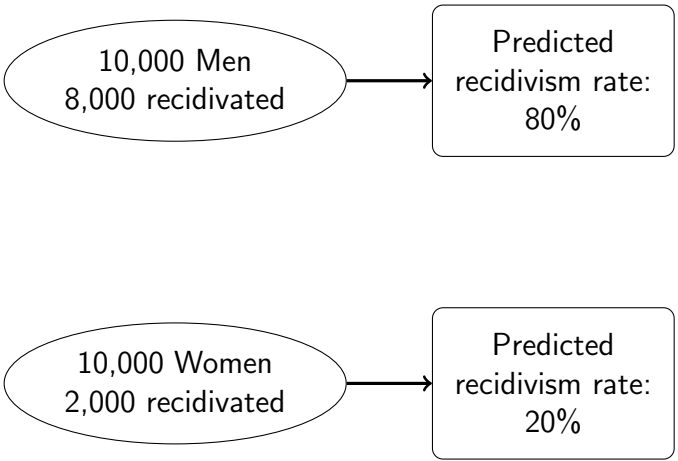
They argued that equal accuracy yielded differences in false positives and false negatives only because the groups have different base rates of recidivism. On this basis, they concluded that COMPAS is fair to Black defendants.

Different False Positive/False Negative Rates

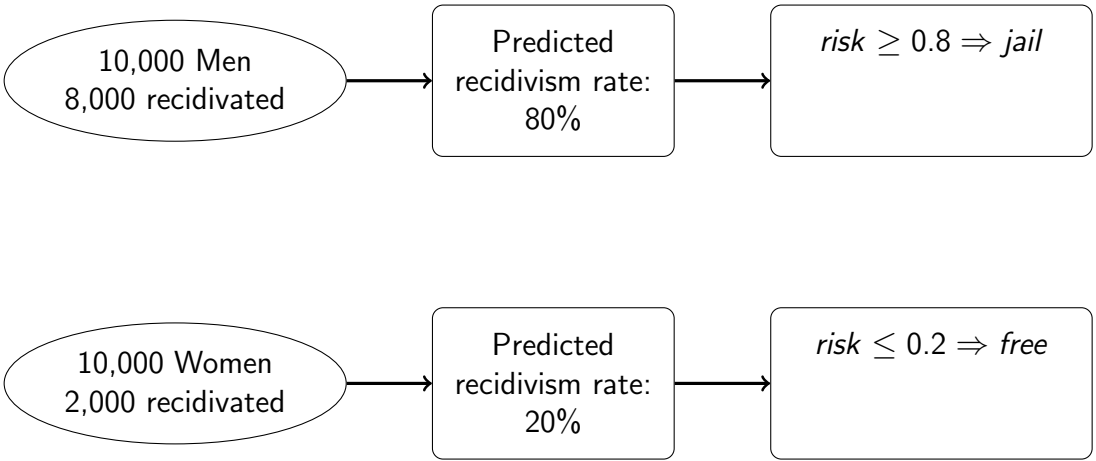




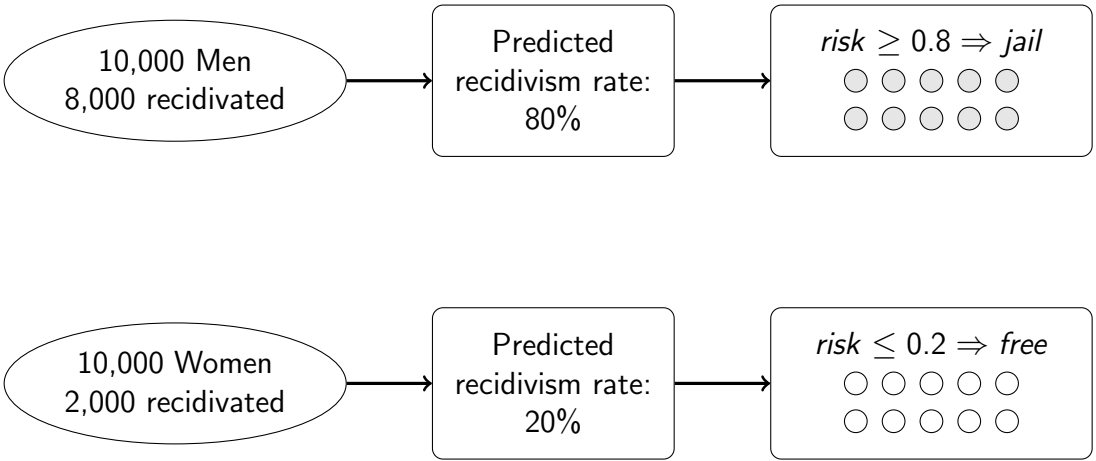
Different False Positive/False Negative Rates



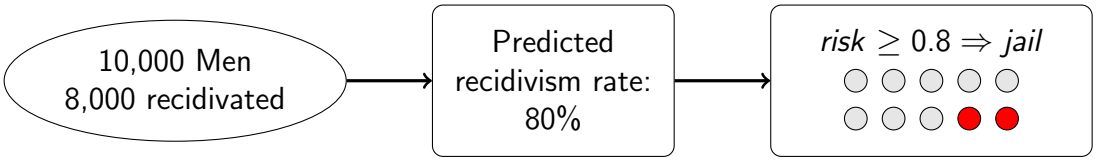
Different False Positive/False Negative Rates



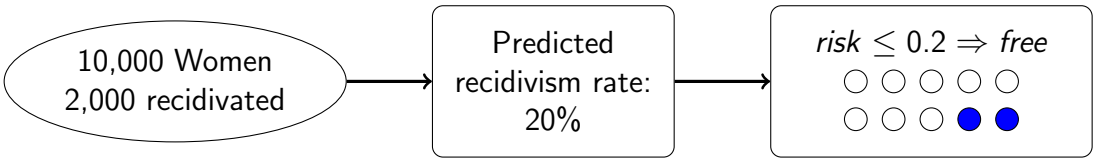
Different False Positive/False Negative Rates



Different False Positive/False Negative Rates

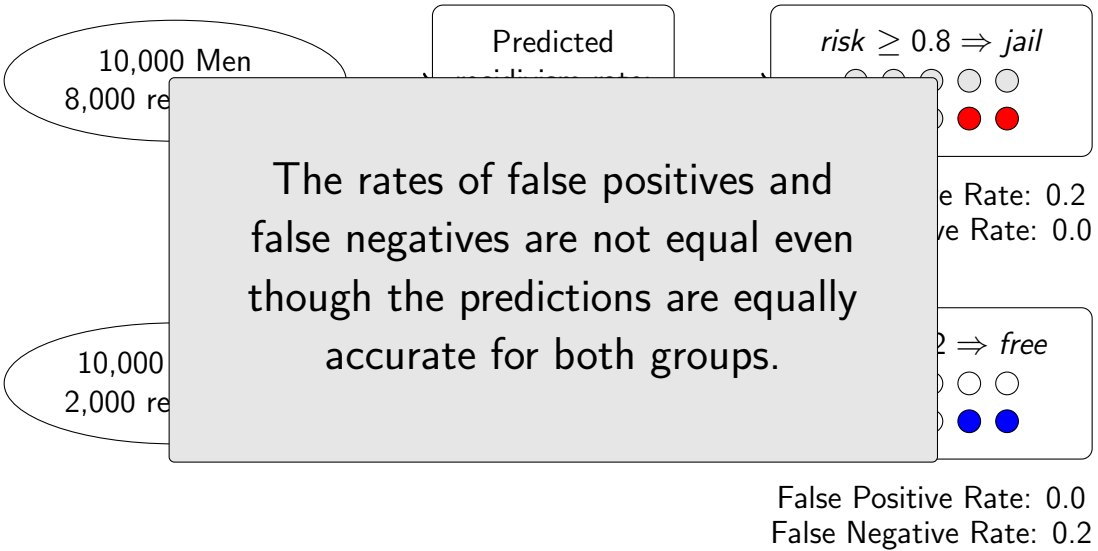


False Positive Rate: 0.2
False Negative Rate: 0.0



False Positive Rate: 0.0
False Negative Rate: 0.2

Different False Positive/False Negative Rates



The rates of false positives and false negatives are not equal even though the predictions are equally accurate for both groups.

Fairness

The issue at stake in these debates concerns which notion of fairness is the right one to guide policy.

- AI is fair when its predictions are equally accurate for different groups.
- AI is fair only when different groups have the same rate of bad outcomes, such as being denied bail, probation, parole, or a shorter sentence.
- AI is fair only when different groups have equal rates of a bad outcome being wrongly imposed, such as bail being denied to those who deserve bail.
- AI is fair when the difference between the average risk scores assigned to the relevant groups should be equal to the difference between the (expected) base rates of those groups.
- ...

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. 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. 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.

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.