PHPE 308M/PHIL 209F
Fairness

Eric Pacuit, University of Maryland

November 10, 2025

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.

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).

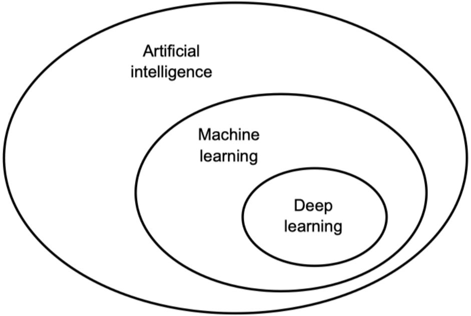
But if AI is so ‘intelligent’, shouldn’t it know better than to be biased?

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 areas where it continues to struggle.

A Brief Introduction to Machine Learning

Machine Learning $\neq$ AI



Learning

The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it...

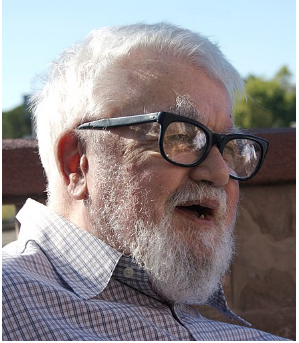
An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.

John McCarthy

Learning

We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

(1952 Dartmouth Workshop)



John McCarthy

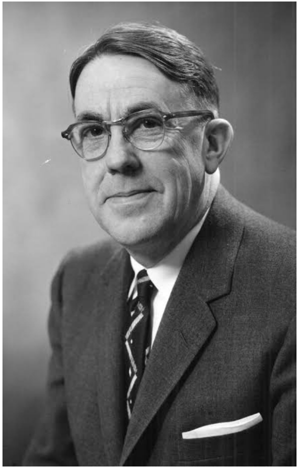
What is Machine Learning?

Learning is any process by which a system improves performance from experience.

Herbert Simon

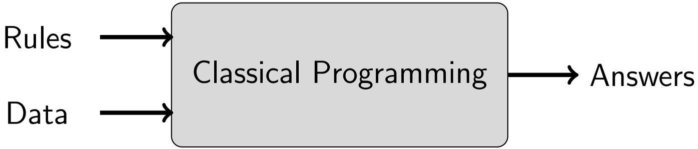
What is Machine Learning?

Machine learning . . . gives computers the ability to learn without being explicitly programmed.

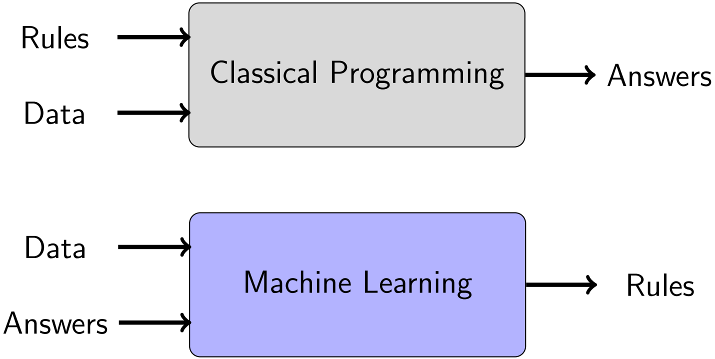


Arthur Lee Samuel

Classical Programming vs. Machine Learning



Classical Programming vs. Machine Learning



Machine Learning

Algorithms that

- improve their **performance** P
- at **task** T
- with **experience** E

A well-defined machine learning task is given by

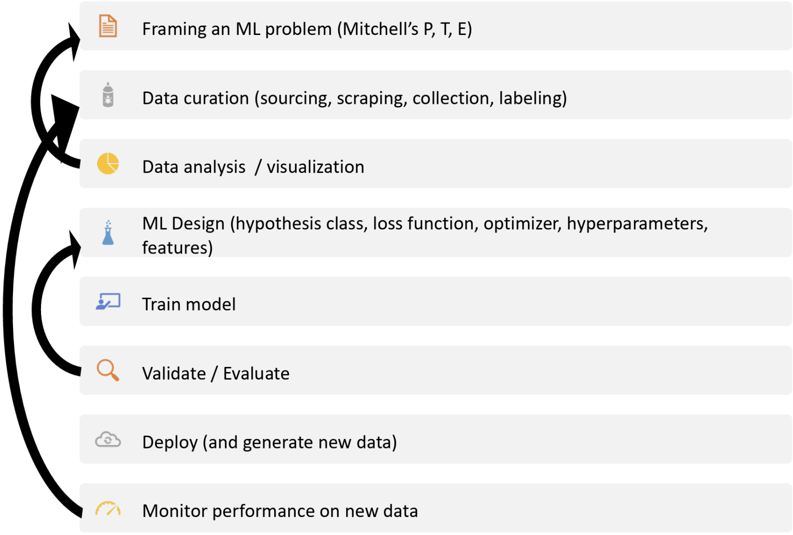
$$(P, T, E).$$

Tom M. Mitchell

Types of Machine Learning Problems

- **Supervised learning**
  - Input: Examples of inputs and outputs
  - Output: Model that predicts unknown output given a new input
- **Unsupervised learning**
  - Input: Examples of some data (no “outputs”)
  - Output: Representation of structure in the data
- **Reinforcement learning**
  - Input: Sequence of interactions with an environment
  - Output: Policy that performs a desired task

Machine Learning Workflow



- Framing an ML problem (Mitchell's P, T, E)
- Data curation (sourcing, scraping, collection, labeling)
- Data analysis / visualization
- ML Design (hypothesis class, loss function, optimizer, hyperparameters, features)
- Train model
- Validate / Evaluate
- Deploy (and generate new data)
- Monitor performance on new data

“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. 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.

- 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.

Example: Determining Bail

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’.

Example: Determining Bail

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.

Example: Determining Bail

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:

1. Will this defendant flee and fail to appear at the trial?
2. Will this defendant commit another crime while out on bail?

Example: Determining 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.

Example: Determining 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.

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

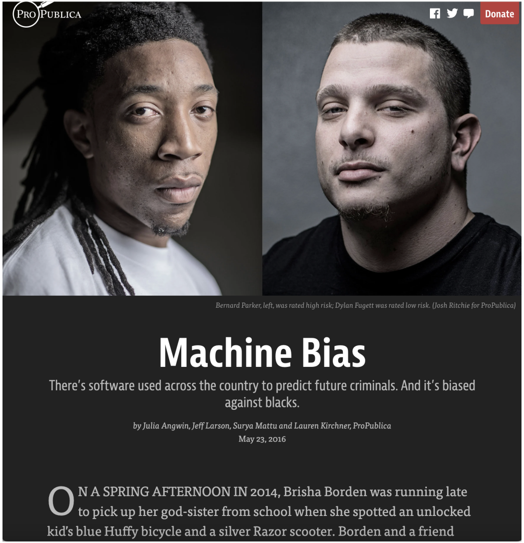
Responsible actors in every sentences system - from prosecutors to judges to parole officials - make daily judgements about....the risks of recidivism posed by offenders. These judgement, pervasive as they are are notoriously imperfect. They often derive from the intuitions and abilities of individual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

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS Algorithm

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

COMPAS Algorithm

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.

COMPAS Algorithm

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.

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.

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

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

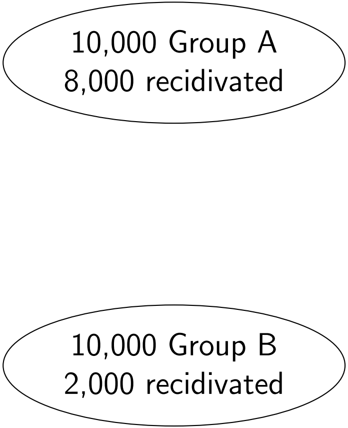
No. Statistical tests isolated the effect of race from criminal history and recidivism, as well as from defendants' age and gender. Black defendants were still 77 per cent more likely to be classified as at higher risk of committing a future violent crime and 45 per cent more likely to be predicted to commit a future crime of any kind.

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.

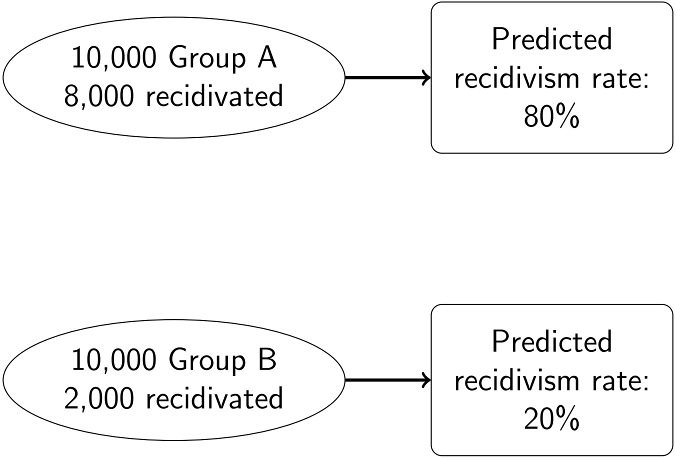
Different False Positive/False Negative Rates



10,000 Group A
8,000 recidivated

10,000 Group B
2,000 recidivated

Different False Positive/False Negative Rates



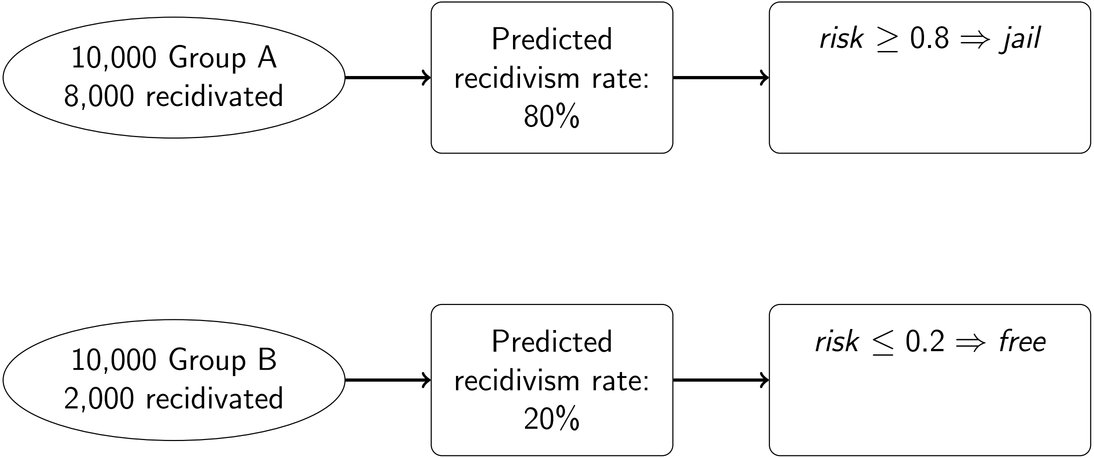
10,000 Group A
8,000 recidivated

Predicted recidivism rate: 80%

10,000 Group B
2,000 recidivated

Predicted recidivism rate: 20%

Different False Positive/False Negative Rates



10,000 Group A
8,000 recidivated

Predicted recidivism rate: 80%

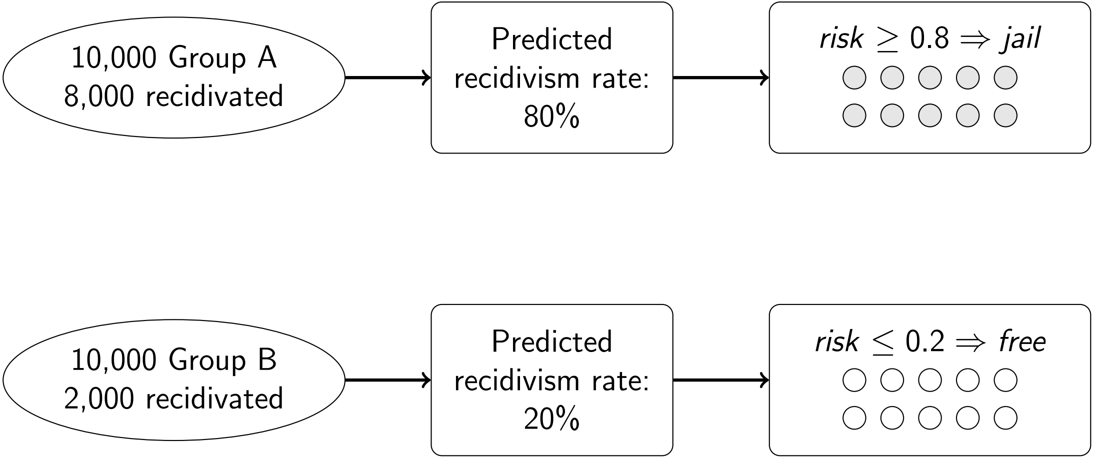
$\mathit{risk} \geq 0.8 \Rightarrow \mathit{jail}$

10,000 Group B
2,000 recidivated

Predicted recidivism rate: 20%

$\mathit{risk} \leq 0.2 \Rightarrow \mathit{free}$

Different False Positive/False Negative Rates



10,000 Group A
8,000 recidivated

Predicted recidivism rate: 80%

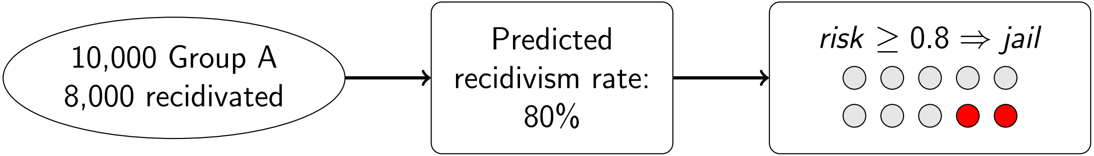
$risk \geq 0.8 \Rightarrow jail$

10,000 Group B
2,000 recidivated

Predicted recidivism rate: 20%

$risk \leq 0.2 \Rightarrow free$

Different False Positive/False Negative Rates

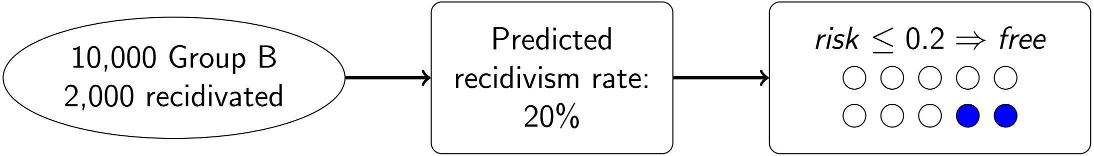


10,000 Group A
8,000 recidivated

Predicted
recidivism rate:
80%

$risk \geq 0.8 \Rightarrow jail$

False Positive Rate: 0.2
False Negative Rate: 0.0



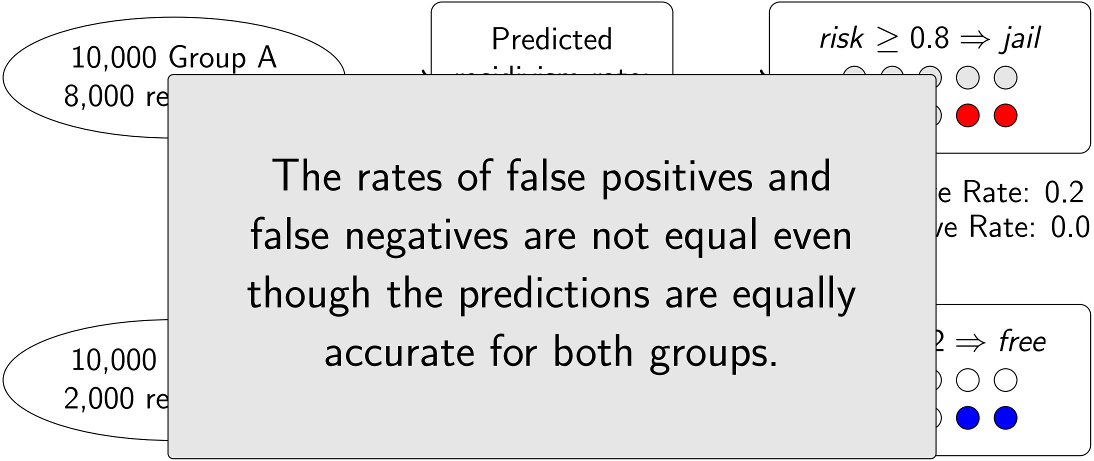
10,000 Group B
2,000 recidivated

Predicted
recidivism rate:
20%

$risk \leq 0.2 \Rightarrow free$

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.

False Positive Rate: 0.0
False Negative Rate: 0.2

Fairness

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

Fairness

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

- AI is fair when its predictions are equally accurate for different groups.

Fairness

The issue at stake 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.

Fairness

The issue at stake 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.

Fairness

The issue at stake 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.

Fairness

The issue at stake 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.