

# PHIL 408Q/PHPE 308D

## Fairness

Eric Pacuit, University of Maryland

April 2, 2024

Kfir Eliaz and Ariel Rubinstein (2014). *On the fairness of random procedures*. Economics Letters, 123, pp. 168 - 170.

After a decision problem and two procedures (Procedure *A* and Procedure *B*) are described. The following is asked to an individual:

In your opinion, from the point of view of (an entity indicated in bold letters):

1. Procedure *A* is fairer than *B* (denoted by *A*)
2. Procedure *B* is fairer than *A* (denoted by *B*) or
3. Both procedures are equally fair (denoted by  $A \sim B$ ).

## P1: randomly pivotal

Consider a committee of 15 members that needs to decide by majority vote whether or not to fire some employee. Simultaneously, each committee member puts his name and his vote in a sealed envelope. The committee chair collects the envelopes and meets in private with the employee. Compare the fairness (from the point of view of the committee members) of the following two procedures for communicating the decision to the employee.



## P1: randomly pivotal

Consider a committee of 15 members that needs to decide by majority vote whether or not to fire some employee. Simultaneously, each committee member puts his name and his vote in a sealed envelope. The committee chair collects the envelopes and meets in private with the employee. Compare the fairness (from the point of view of the committee members) of the following two procedures for communicating the decision to the employee.

- (A) The committee chair opens the envelopes in private and counts the votes. He announces the outcome of the vote to the candidate and shows him the content of each envelope in some random order.

## P1: randomly pivotal

Consider a committee of 15 members that needs to decide by majority vote whether or not to fire some employee. Simultaneously, each committee member puts his name and his vote in a sealed envelope. The committee chair collects the envelopes and meets in private with the employee. Compare the fairness (from the point of view of the committee members) of the following two procedures for communicating the decision to the employee.

- (A) The committee chair opens the envelopes in private and counts the votes. He announces the outcome of the vote to the candidate and shows him the content of each envelope in some random order.
- (B) The committee chair opens the envelopes in some random order in front of the candidate. For each opened envelope he announces the name of the committee member and his vote. When at some point, a majority of votes is reached the chair announces the outcome and continues to open the remaining envelopes.

## P1: randomly pivotal

Procedure  $A$  is intuitively fairer than  $B$  since in  $B$  one of the committee members appears to be responsible for the firing decision, in violation of:

(C1) It is fair to treat all individuals equally ex-ante.

## P1: randomly pivotal

Procedure  $A$  is intuitively fairer than  $B$  since in  $B$  one of the committee members appears to be responsible for the firing decision, in violation of:

(C1) It is fair to treat all individuals equally ex-ante.

Results:

$A$	$B$	$A \sim B$
56%	18%	26%

## P2: random dictatorship

You are a student in a class that needs to select one of two exam dates.  
Compare the fairness (from the point of view of the students) of the following procedures for making the decision.

## P2: random dictatorship

You are a student in a class that needs to select one of two exam dates. Compare the fairness (from the point of view of the students) of the following procedures for making the decision.

- (A) One of the students is selected at random and is asked to make the choice. His identity will be announced and his decision will determine the outcome.
- (B) Each student has to submit a note bearing his name and his choice. One of the notes will be randomly picked; the identity of the student will be announced and his choice will determine the outcome.

## P2: random dictatorship

The two procedures are versions of the “random dictator” voting method. Both treat all individuals equally ex-ante (it satisfies (C1)), but only Procedure *B* is more likely to be viewed as fairer since it is the only one satisfying:

- (C2) It is fair to allow all individuals to actively participate in the procedure whatever the realization of the random elements.

## P2: random dictatorship

The two procedures are versions of the “random dictator” voting method. Both treat all individuals equally ex-ante (it satisfies (C1)), but only Procedure  $B$  is more likely to be viewed as fairer since it is the only one satisfying:

- (C2) It is fair to allow all individuals to actively participate in the procedure whatever the realization of the random elements.

Results:

$A$	$B$	$A \sim B$
5%	52%	43%



## P3: implicit or explicit randomization

Consider an employer who needs to fire at most one worker who failed some qualification exam. All workers have taken the exam, some passed some failed. Compare the fairness (from the point of view of the workers) of the following procedures for selecting the worker to be fired.

### P3: implicit or explicit randomization

Consider an employer who needs to fire at most one worker who failed some qualification exam. All workers have taken the exam, some passed some failed. Compare the fairness (from the point of view of the workers) of the following procedures for selecting the worker to be fired.

- (A) The employer reviews the list of exam results at a random order. The first worker to fail the exam is fired.
- (B) The employer selects a worker at random from among all the workers who failed the exam.

This problem is related to experiment 9 in Keren and Teigen (2010). They asked subjects to rank four types of random procedures for deciding which patient will receive treatment. Their findings indicate a tendency to view a coin toss as fairer than procedures such as drawing a piece of paper out of a hat or randomly choosing one of the rooms in which each patient is waiting.

Gideon Keren and Karl H. Teigen (2010). *Decisions by coin toss: Inappropriate but fair*. Judgment and Decision Making, 5(2), pp. 83 - 101.

## P3: implicit or explicit randomization

Both procedures satisfy (C1) and (C2): Ex ante, each worker who failed the exam has the same chance of being fired. In addition, all workers actively participate in the procedure by taking the exam.

### P3: implicit or explicit randomization

Both procedures satisfy (C1) and (C2): Ex ante, each worker who failed the exam has the same chance of being fired. In addition, all workers actively participate in the procedure by taking the exam.

Both procedures have two stages: In  $A$ , the random element is activated first and then the exams are marked; In  $B$ , all exams are marked and then the random element is realized. But only  $B$  satisfies the following:

- (C3) It is fair to delay any asymmetry in the treatment of participants to as late a stage as possible in the procedure.

### P3: implicit or explicit randomization

Both procedures satisfy (C1) and (C2): Ex ante, each worker who failed the exam has the same chance of being fired. In addition, all workers actively participate in the procedure by taking the exam.

Both procedures have two stages: In  $A$ , the random element is activated first and then the exams are marked; In  $B$ , all exams are marked and then the random element is realized. But only  $B$  satisfies the following:

- (C3) It is fair to delay any asymmetry in the treatment of participants to as late a stage as possible in the procedure.

Results:

$A$	$B$	$A \sim B$
6%	40%	54%

## P4: the doctor or the mother

Suppose two twins need to receive a kidney transplant from their mother. The mother can donate only one kidney. Compare the fairness (from the point of view of the mother) of the following two procedures for determining who will receive the kidney.

## P4: the doctor or the mother

Suppose two twins need to receive a kidney transplant from their mother. The mother can donate only one kidney. Compare the fairness (from the point of view of the mother) of the following two procedures for determining who will receive the kidney.

- (A) The doctor will toss a coin.
- (B) The mother will toss the coin.



## P4: the doctor or the mother

If the mother tosses the coin, she will bear a higher psychological burden than the doctor as a result of denying a kidney to one of her children. Only *A* satisfies the following:

- (C4) It is fair to reduce the psychological burden associated with the perception that the individual who executes a random device bears some responsibility for its outcome.

## P4: the doctor or the mother

If the mother tosses the coin, she will bear a higher psychological burden than the doctor as a result of denying a kidney to one of her children. Only *A* satisfies the following:

- (C4) It is fair to reduce the psychological burden associated with the perception that the individual who executes a random device bears some responsibility for its outcome.

Results:

<i>A</i>	<i>B</i>	$A \sim B$
31%	10%	58%

## P5: the 'drawn' or the 'not drawn'

Imagine there are two equally qualified candidates for a position, both of whom reached the final stage of the recruiting process. The name of each candidate is put in a sealed envelope. One of the envelopes will be randomly drawn. Compare the fairness (from the point of view of the candidates) of the following two procedures for selecting the candidate to be hired.

## P5: the 'drawn' or the 'not drawn'

Imagine there are two equally qualified candidates for a position, both of whom reached the final stage of the recruiting process. The name of each candidate is put in a sealed envelope. One of the envelopes will be randomly drawn. Compare the fairness (from the point of view of the candidates) of the following two procedures for selecting the candidate to be hired.

- (A) The candidate whose name is drawn is hired.
- (B) The candidate whose name is not drawn is hired.

## P5: the 'drawn' or the 'not drawn'

A appears to be fairer according to two fairness criteria.

(C5) It is fair to use “conventional” or “familiar” means of randomization.

(C6) It is fair to respect “divine providence” as manifested  
in the realization of the random device.

## P5: the ‘drawn’ or the ‘not drawn’

$A$  appears to be fairer according to two fairness criteria.

(C5) It is fair to use “conventional” or “familiar” means of randomization.

(C6) It is fair to respect “divine providence” as manifested  
in the realization of the random device.

Results:

$A$	$B$	$A \sim B$
14%	2%	84%

## P6: drawn twice

One prize is to be awarded to one person from among 20 candidates. Compare the fairness (from the point of view of the candidates) of the following procedures for selecting who will get the prize.

## P6: drawn twice

One prize is to be awarded to one person from among 20 candidates. Compare the fairness (from the point of view of the candidates) of the following procedures for selecting who will get the prize.

- (A) A computer program repeatedly draws a name at random, and the prize is awarded to the first person whose name is drawn twice.
- (B) A computer program draws one of the names at random and that person is awarded the prize.



## P6: drawn twice

There are two conflicting criteria in this case.

On the one hand, the fact that the same name appears twice is an indication that it is “God’s will” and thus according to (C6) procedure *A* is fairer.

On the other hand, Procedure *A* allows for candidates to be drawn once but not to be selected in the end, which may be viewed as going against “God’s will” and thus, according to (C6) Procedure *B* is fairer.

## P6: drawn twice

There are two conflicting criteria in this case.

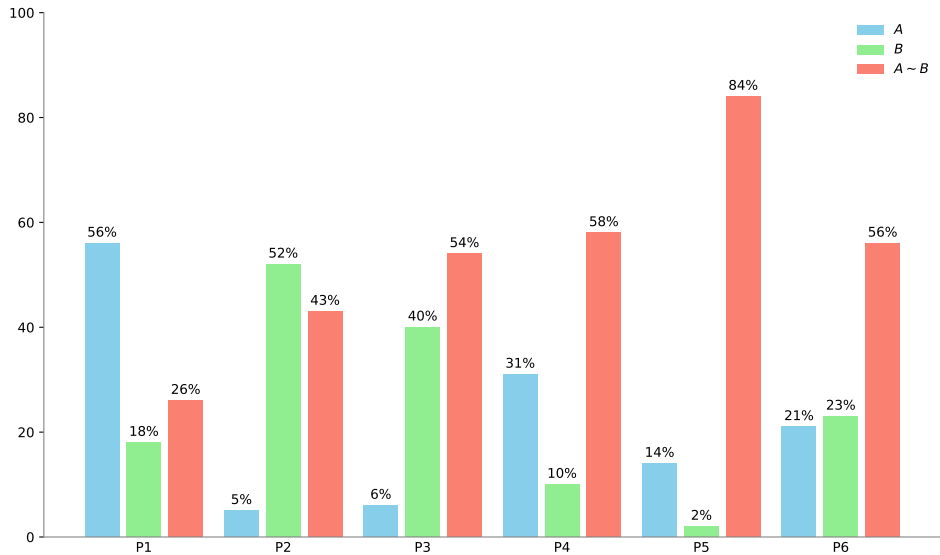
On the one hand, the fact that the same name appears twice is an indication that it is “God’s will” and thus according to (C6) procedure *A* is fairer.

On the other hand, Procedure *A* allows for candidates to be drawn once but not to be selected in the end, which may be viewed as going against “God’s will” and thus, according to (C6) Procedure *B* is fairer.

Results:

<i>A</i>	<i>B</i>	$A \sim B$
21%	23%	56%

# Results



- (C1) It is fair to treat all individuals equally ex-ante.
- (C2) It is fair to allow all individuals to actively participate in the. procedure whatever the realization of the random elements.
- (C3) It is fair to delay any asymmetry in the treatment of participants to as late a stage as possible in the procedure.
- (C4) It is fair to reduce the psychological burden associated with the perception that the individual who executes a random device bears some responsibility for its outcome.
- (C5) It is fair to use “conventional” or “familiar” means of randomization.
- (C6) It is fair to respect “divine providence” as manifested in the realization of the random device.

A natural question is whether the data points to the existence of “types”, i.e., systematic patterns in responses that characterize significant proportions of the participants.

The proposed typology is based on only the first four questions. This is because 84% of the subjects in P5 considered both procedures to be equally fair and no unique procedure was perceived as being fairer than the other in P6.

# Types of Responses

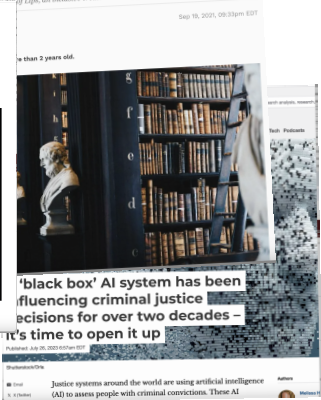
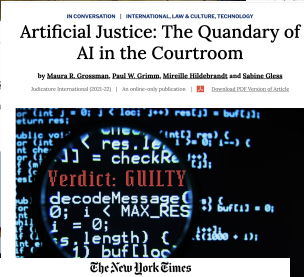
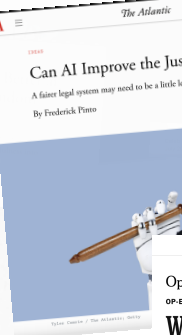
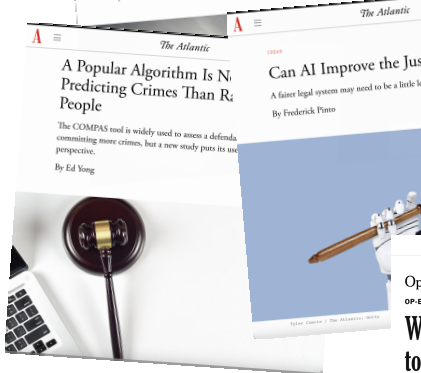
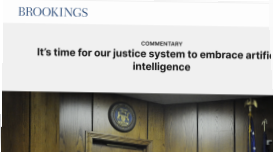
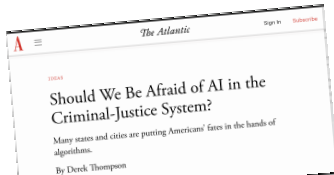
Emotional			
P1	P2	P3	P4
A	B	B	A
A	B	B	B
A	B	B	$A \sim B$
A	B	A	A
A	B	$A \sim B$	A
B	B	B	A
$A \sim B$	B	B	A

Consequentialist			
P1	P2	P3	P4
$A \sim B$	$A \sim B$	$A \sim B$	$A \sim B$
$A \sim B$	$A \sim B$	$A \sim B$	A
$A \sim B$	$A \sim B$	$A \sim B$	B
$A \sim B$	$A \sim B$	A	$A \sim B$
$A \sim B$	$A \sim B$	B	$A \sim B$
$A \sim B$	A	$A \sim B$	$A \sim B$
$A \sim B$	B	$A \sim B$	$A \sim B$
A	$A \sim B$	$A \sim B$	$A \sim B$
B	$A \sim B$	$A \sim B$	$A \sim B$

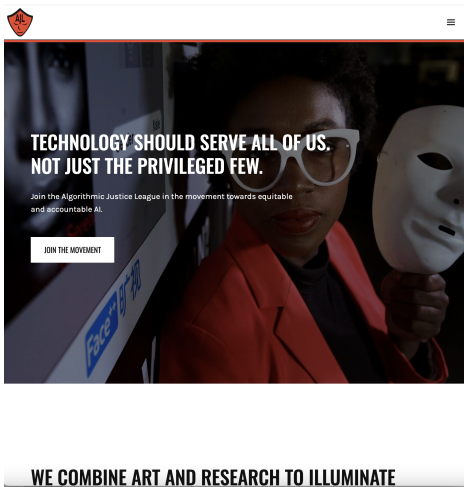
- ▶ Consequentialist: About 31% of the subjects fall into this category. Of those 209 subjects, 40% displayed four indifferences and 60% displayed three.
- ▶ Emotional: About 30% of all participants were classified as emotional and 25% of them chose exactly  $(A, B, B, A)$ .
- ▶ Other: Chosen by only 39% of the subjects. Each of these profiles was exhibited by at most 6% of all subjects.

# Fairness in AI





# Algorithmic Justice League



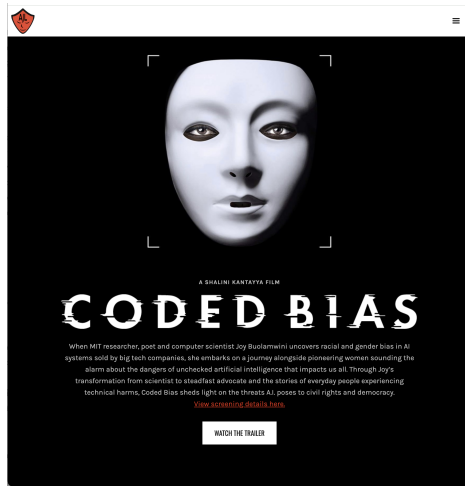
The banner features a woman with glasses and a red jacket holding a white mask. In the background, a 'Face++' logo is visible. The text is overlaid on the left side of the image.

**TECHNOLOGY SHOULD SERVE ALL OF US.  
NOT JUST THE PRIVILEGED FEW.**

Join the Algorithmic Justice League in the movement towards equitable and accountable AI.

[JOIN THE MOVEMENT](#)

**WE COMBINE ART AND RESEARCH TO ILLUMINATE**



The poster features a white mask with a face, set against a black background. The mask is framed by a white border. The text is centered below the mask.

A SHALINI KANTAYYA FILM

**CODED BIAS**

When MIT researcher, poet and computer scientist Joy Buolamwini uncovers racial and gender bias in AI systems sold by big tech companies, she embarks on a journey alongside pioneering women sounding the alarm about the dangers of unchecked artificial intelligence that impacts us all. Through Joy's transformation from scientist to steadfast advocate and the stories of everyday people experiencing technical harms, Coded Bias sheds light on the threats AI poses to civil rights and democracy.

[View screening details here.](#)

[WATCH THE TRAILER](#)

# AI and Fairness @ UMD



## Welcome to VCAI!

The goal of Values-Centered Artificial Intelligence (VCAI) is to integrate research and education across campus, engage in high-impact research with local stakeholders, and – hopefully – transform how artificial intelligence is practiced globally.

---

How can you get involved?

[VCAI Mailing List](#)

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 Als end up biased is that humans and human social structures are often biased, and our biases are readily built into the Als 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.