Biases in AI

Bias vs discrimination

Bias:

- A prejudiced attitude or preconceived notion about a group
- An internal mental tendency or predisposition
- Can exist without direct action
- Often unconscious or unintentional
- Represents personal beliefs or stereotypes

Discrimination:

- Actual harmful actions or treatment based on bias
- Concrete behaviors that exclude or disadvantage a group
- Tangible and observable differential treatment
- Involves denying opportunities, rights, or privileges
- Typically involves systematic or institutional practices

A utilitarian point of view

Utilitarian arguments against bias:

- Biases create economic inefficiencies by preventing optimal talent deployment
- Discriminatory practices reduce aggregate happiness
- Biases generate psychological harm and social tension
- Eliminating biases could maximize overall societal utility

Utilitarian Solution Approach:

- Systematically identify and mitigate biases
- Design institutional structures that minimize biased outcomes
- Create incentive structures that reward unbiased decision-making
- Develop measurement tools to quantify bias's negative social impacts

The core utilitarian concern would be: ***How do biases reduce net social welfare, and what interventions could mo***
effectively increase overall human well-being?

A Kantian's view on bias

Key Kantian critiques of bias:

- Biases reduce individuals to stereotypical categories, undermining their inherent human dignity
- They contradict the categorical imperative of universal moral law, which demands treating all rational beings with equal respect
- Biases represent a failure of rational thought, prioritizing prejudiced, non-universal thinking over reasoned judgment

Kant would argue that genuine moral reasoning requires:

- Recognizing each person's intrinsic worth
- Developing rational, impartial perspectives
- Transcending subjective, prejudiced viewpoints
- Applying universal ethical principles consistently

In Kant's framework, biases are not just morally wrong but fundamentally irrational - they represent a collapse of reasoned, principled thinking into arbitrary, subjective categorizations that deny individuals their autonomy.

Virtue ethics and biases

Biases are character flaws that prevent the development of practical wisdom and moral excellence.

Key Virtues Challenged by Biases:

- Fairness (ability to judge impartially)
- Magnanimity (generosity of spirit)
- Prudence (rational decision-making)

Aristotelian Analysis:

- Biases represent a deviation from the "golden mean" between extremes
- They indicate a failure of character development
- Reflect an inability to cultivate rational, balanced judgment

Virtue ethics and biases

Virtuous Response to Biases:

- Continuous self-reflection
- Developing empathy and understanding
- Practicing intellectual humility
- Actively challenging one's preconceptions
- Cultivating a more nuanced, compassionate worldview

The goal would be to transform biased thinking through habitual practice of virtuous reasoning, ultimately achieving a more balanced, morally refined character that sees the inherent worth in all individuals.

Is it wrong to have biases?

The existence of implicit biases is not inherently "wrong" - they're a natural cognitive process resulting from our brain's pattern-recognition mechanisms. *What matters ethically is:*

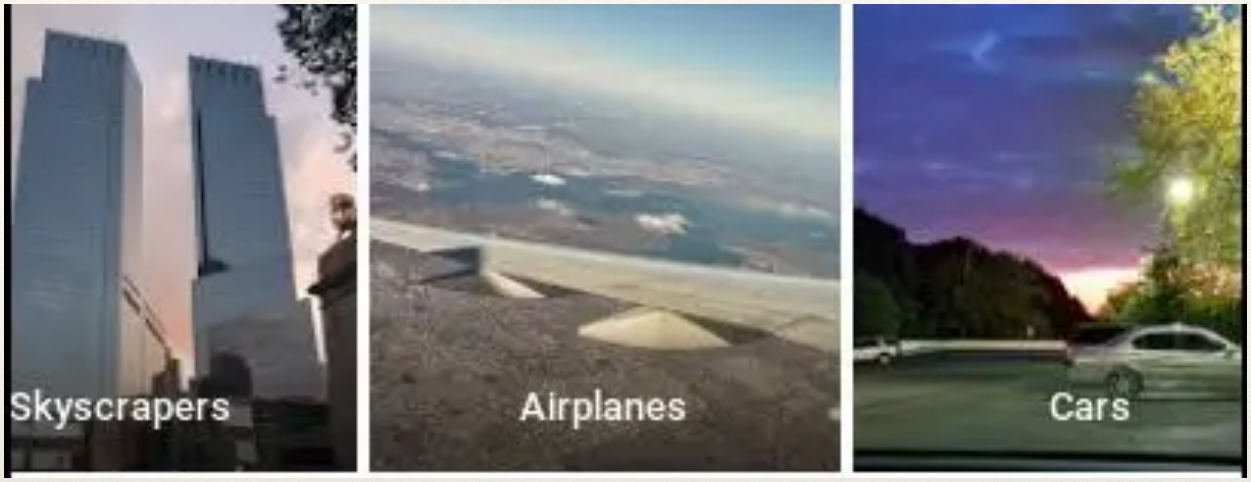
- Awareness of these biases
- Active efforts to recognize and mitigate them
- Preventing biases from translating into discriminatory behaviors

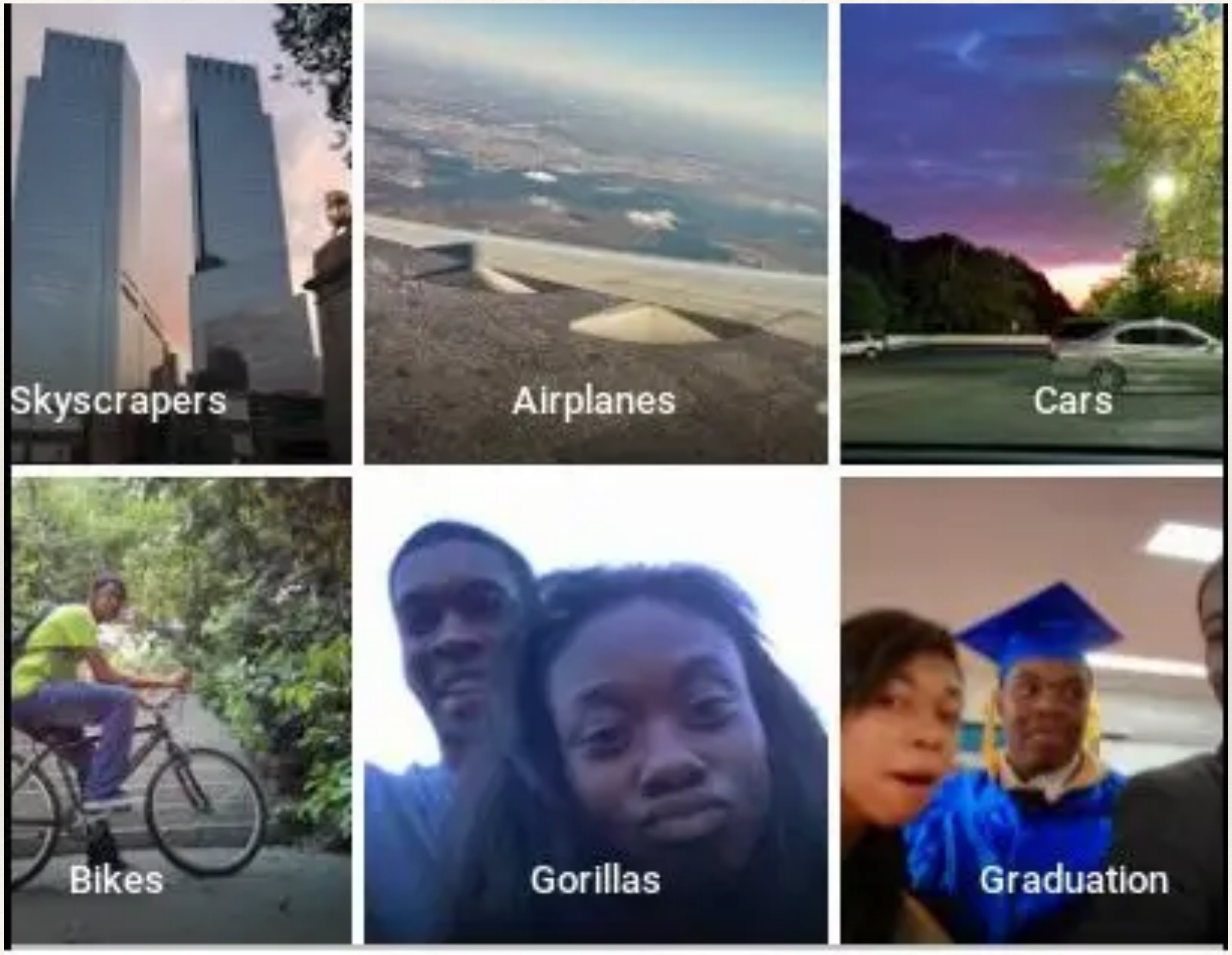
Key points:

- Implicit biases are unconscious and universal
- They don't automatically make someone prejudiced
- The moral responsibility lies in how we respond to these biases

The ethical approach is not to feel guilty about having implicit biases, but to:

- Acknowledge their existence
- Develop strategies to minimize their impact
- Create systems that reduce bias-driven decision-making
- Cultivate empathy and understanding


Skyscrapers
Airplanes
Cars



https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html

Claude’s point of view

Ethical Implications:

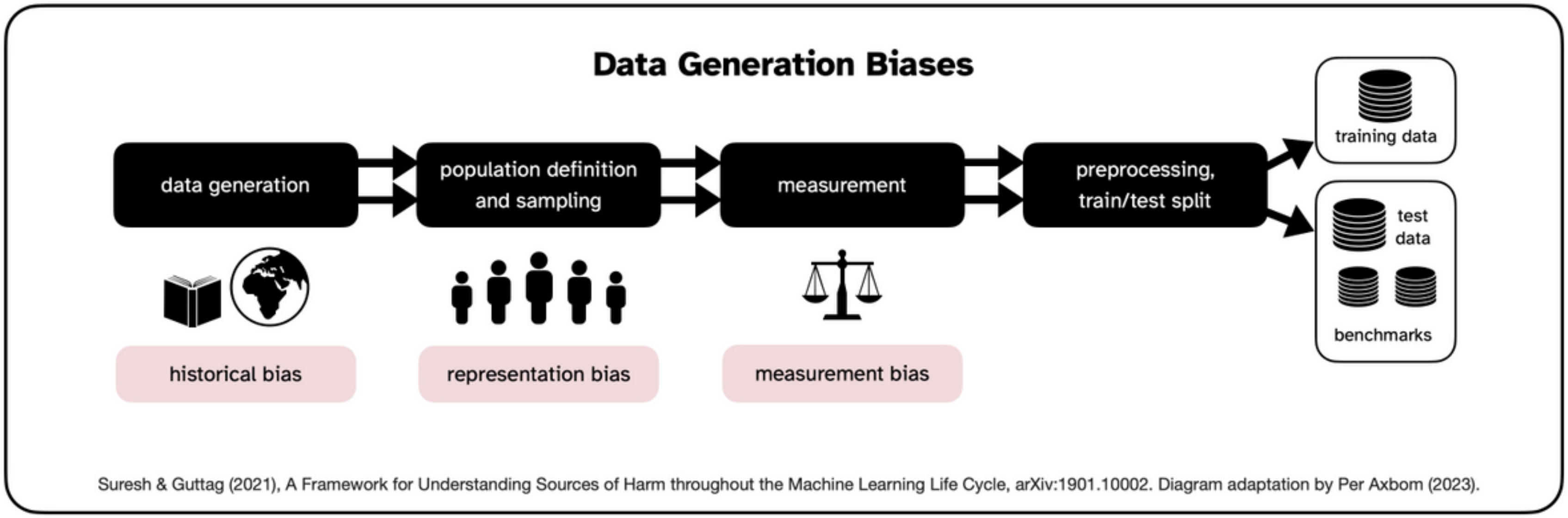
- Perpetuates harmful racial stereotypes
- Demonstrates systemic racism embedded in technological systems
- Causes psychological harm and reinforces dehumanizing narratives

Moral Responsibility:

- Technology companies must proactively address algorithmic bias
- Requires diverse teams and rigorous ethical review processes
- Necessitates ongoing auditing of recognition technologies

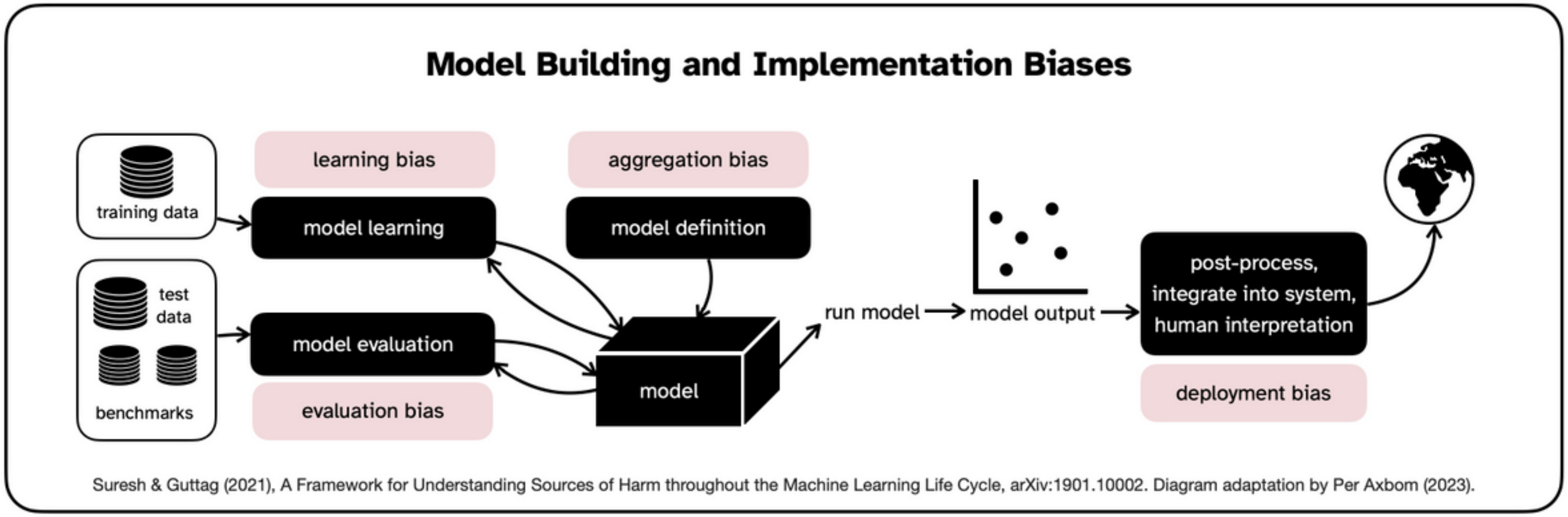
Broader Context:

- Reflects deeper societal issues of racial discrimination
- Highlights the need for inclusive technological development
- Demonstrates how unexamined biases can be inadvertently reproduced through technology



Data Generation Biases

Suresh & Guttag (2021), A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, arXiv:1901.10002. Diagram adaptation by Per Axbom (2023).



Model Building and Implementation Biases

Suresh & Guttag (2021), A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, arXiv:1901.10002. Diagram adaptation by Per Axbom (2023).

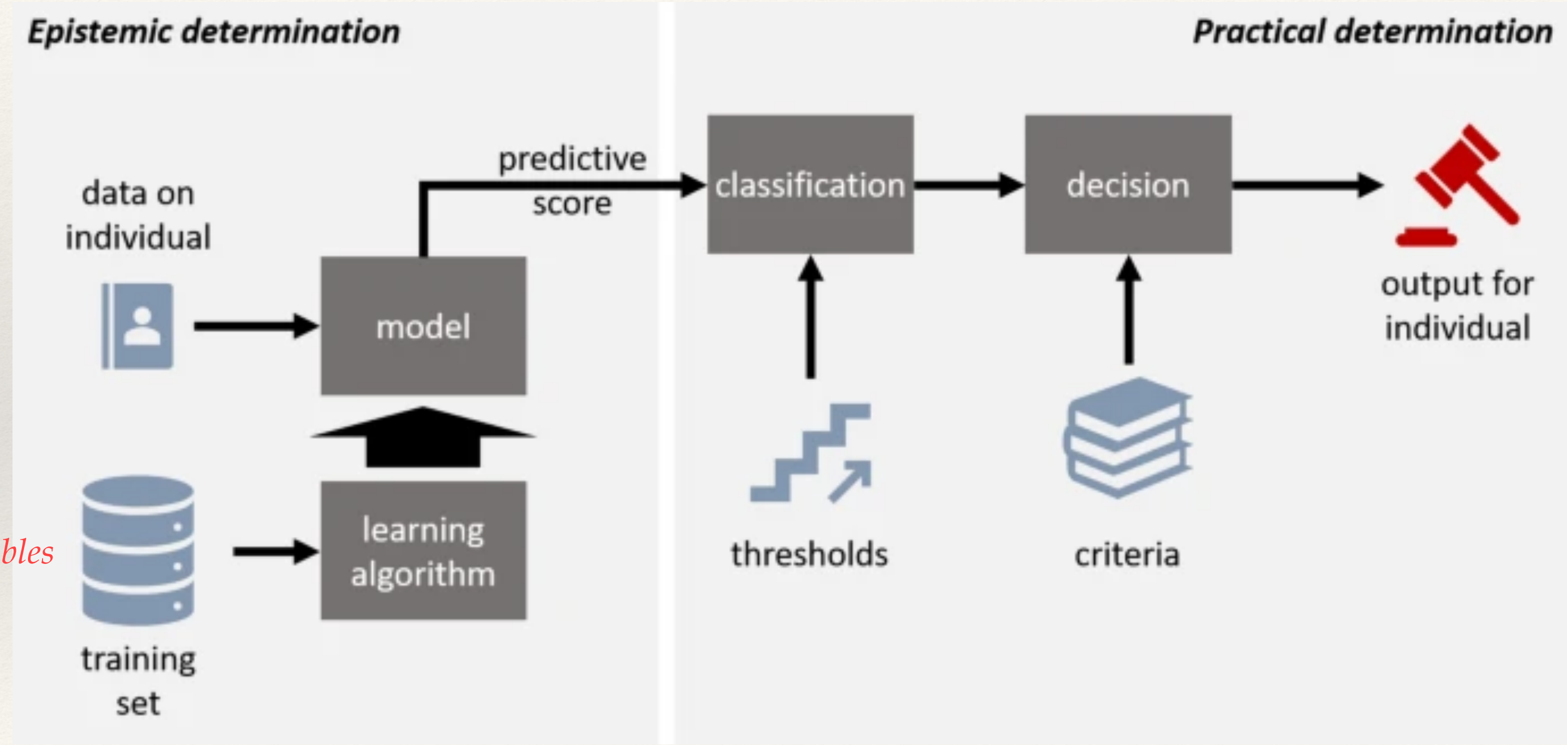
COMPAS

Correctional Offender Management Profiling for Alternative Sanctions

Designed by NorthPointe

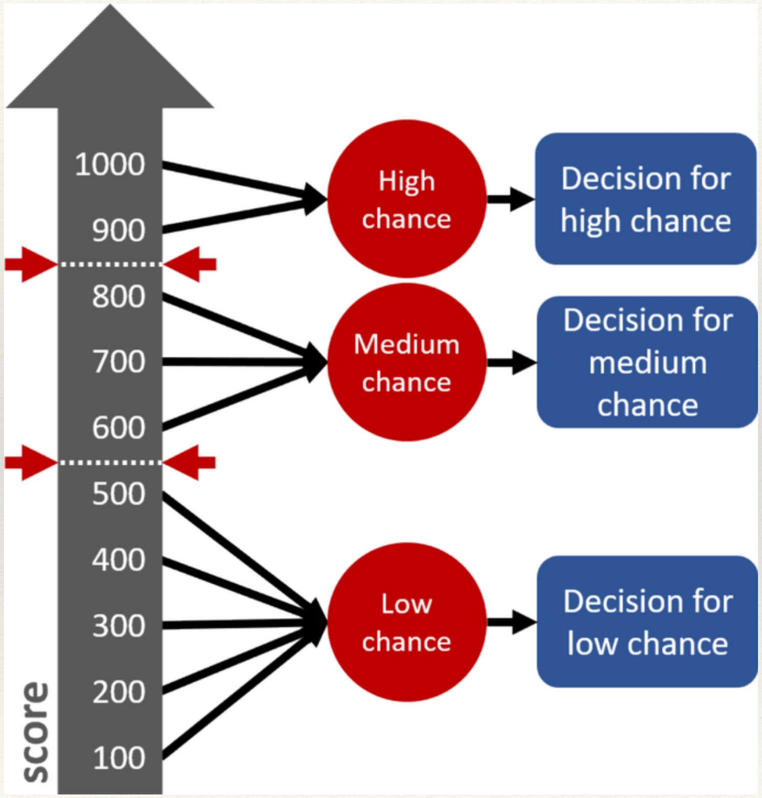
COMPAS is a risk/needs assessment instrument used by criminal justice agencies to make informed decisions regarding placement, supervision, and case management of offenders.

COMPAS

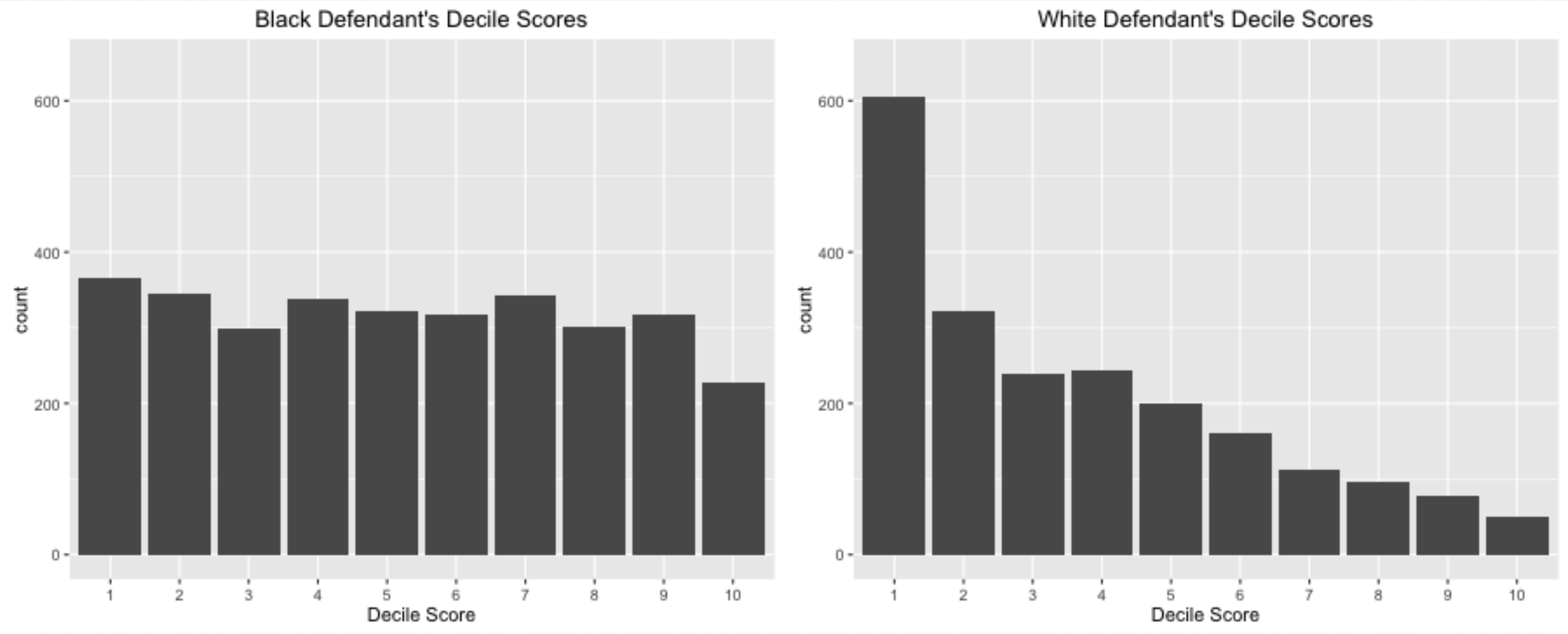


137 variables

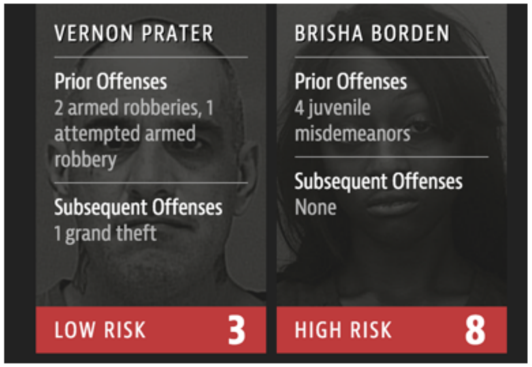
COMPAS


1000
900
800
700
600
500
400
300
200
100
score
High chance
Medium chance
Low chance
Decision for high chance
Decision for medium chance
Decision for low chance

COMPAS


Black Defendant's Decile Scores
White Defendant's Decile Scores
count
600
400
200
0
1
2
3
4
5
6
7
8
9
10
Decile Score

Race biases



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

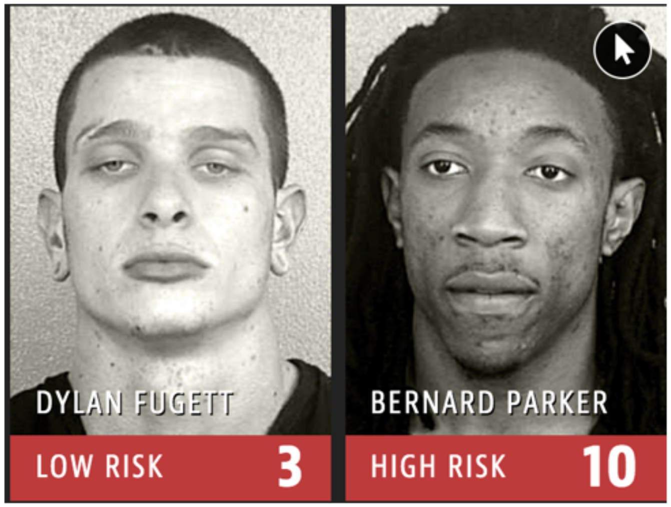
LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

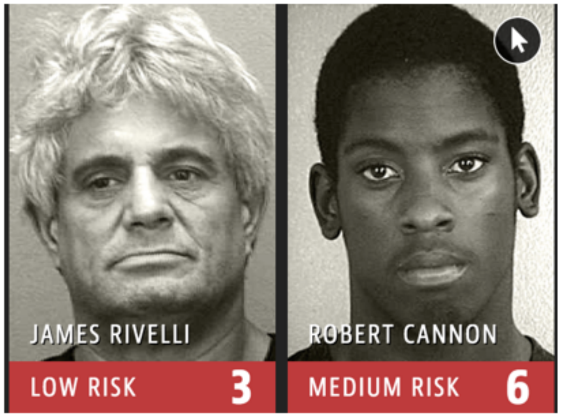


DYLAN FUGETT

LOW RISK 3

BERNARD PARKER

HIGH RISK 10

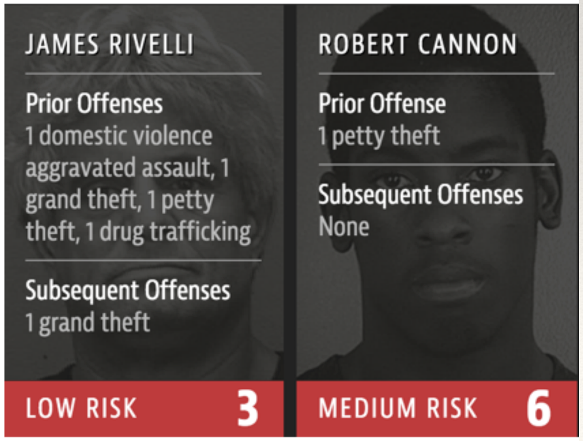


JAMES RIVELLI

LOW RISK 3

ROBERT CANNON

MEDIUM RISK 6



JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

LOW RISK 3

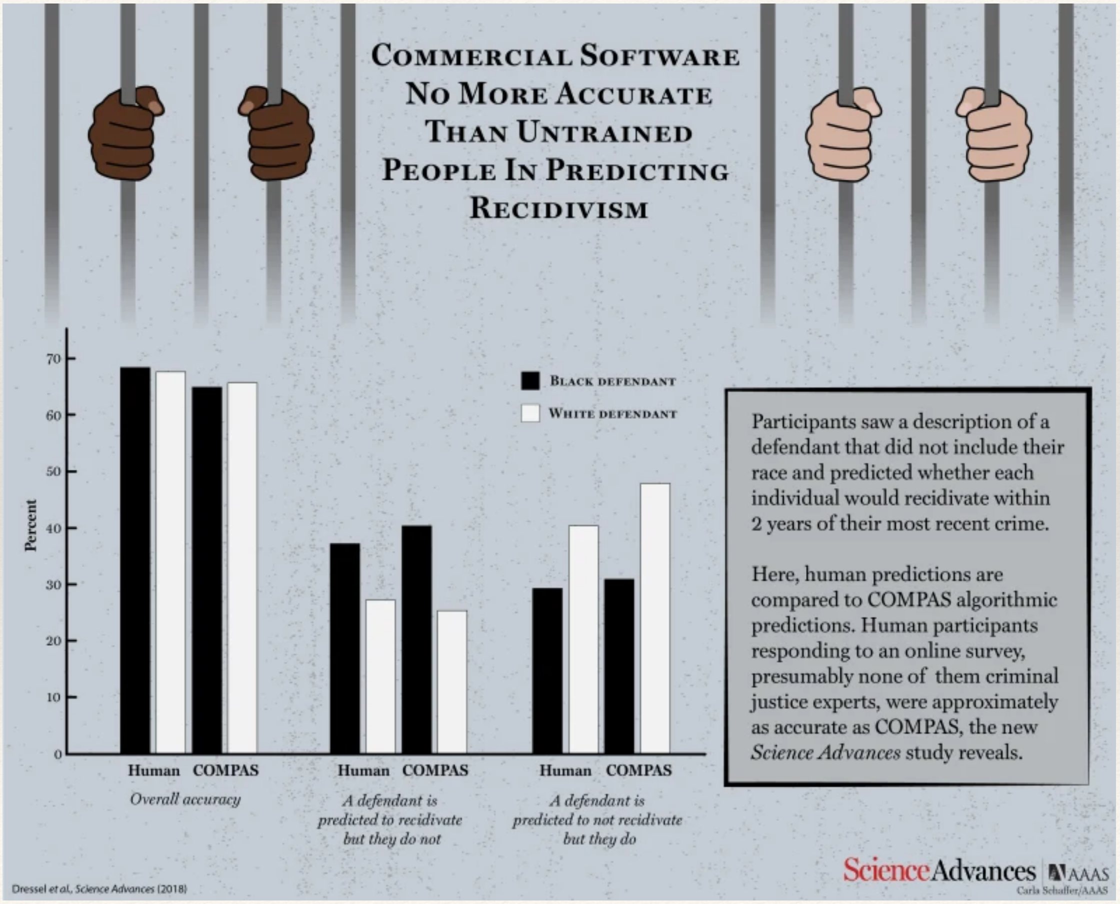
ROBERT CANNON

Prior Offense
1 petty theft

Subsequent Offenses
None

MEDIUM RISK 6

COMPAS



Commercial Software No More Accurate Than Untrained People In Predicting Recidivism

Participants saw a description of a defendant that did not include their race and predicted whether each individual would recidivate within 2 years of their most recent crime.

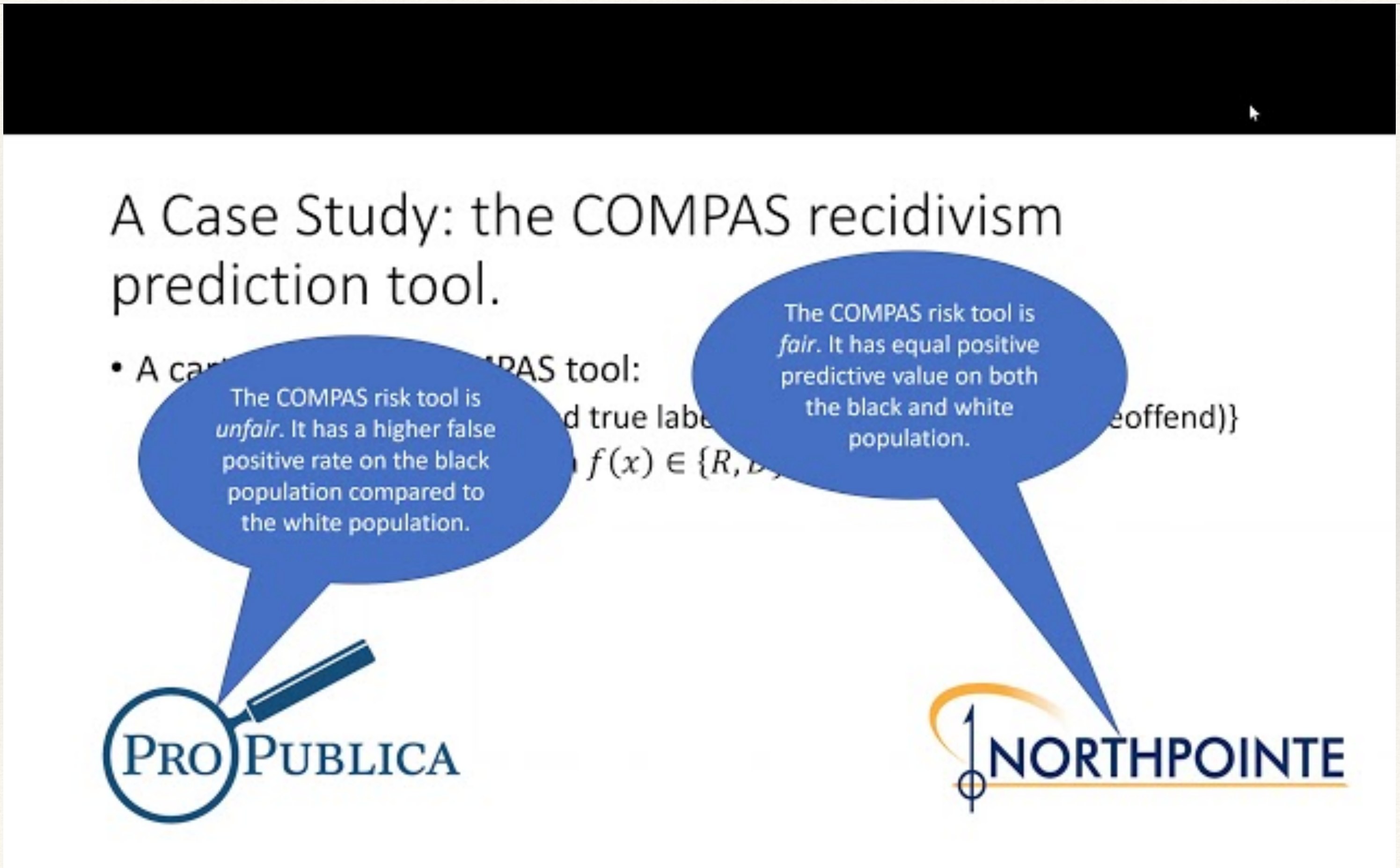
Here, human predictions are compared to COMPAS algorithmic predictions. Human participants responding to an online survey, presumably none of them criminal justice experts, were approximately as accurate as COMPAS, the new *Science Advances* study reveals.

Dressel et al., Science Advances (2018)

ScienceAdvances | AAAS

Carla Schaffer/AAAS

COMPAS



A Case Study: the COMPAS recidivism prediction tool.

- A ca PAS tool:

d true labe eoffend)}

$f(x) \in \{R, D$

The COMPAS risk tool is *unfair*. It has a higher false positive rate on the black population compared to the white population.

The COMPAS risk tool is *fair*. It has equal positive predictive value on both the black and white population.

PROPUBLICA

NORTHPOINTE

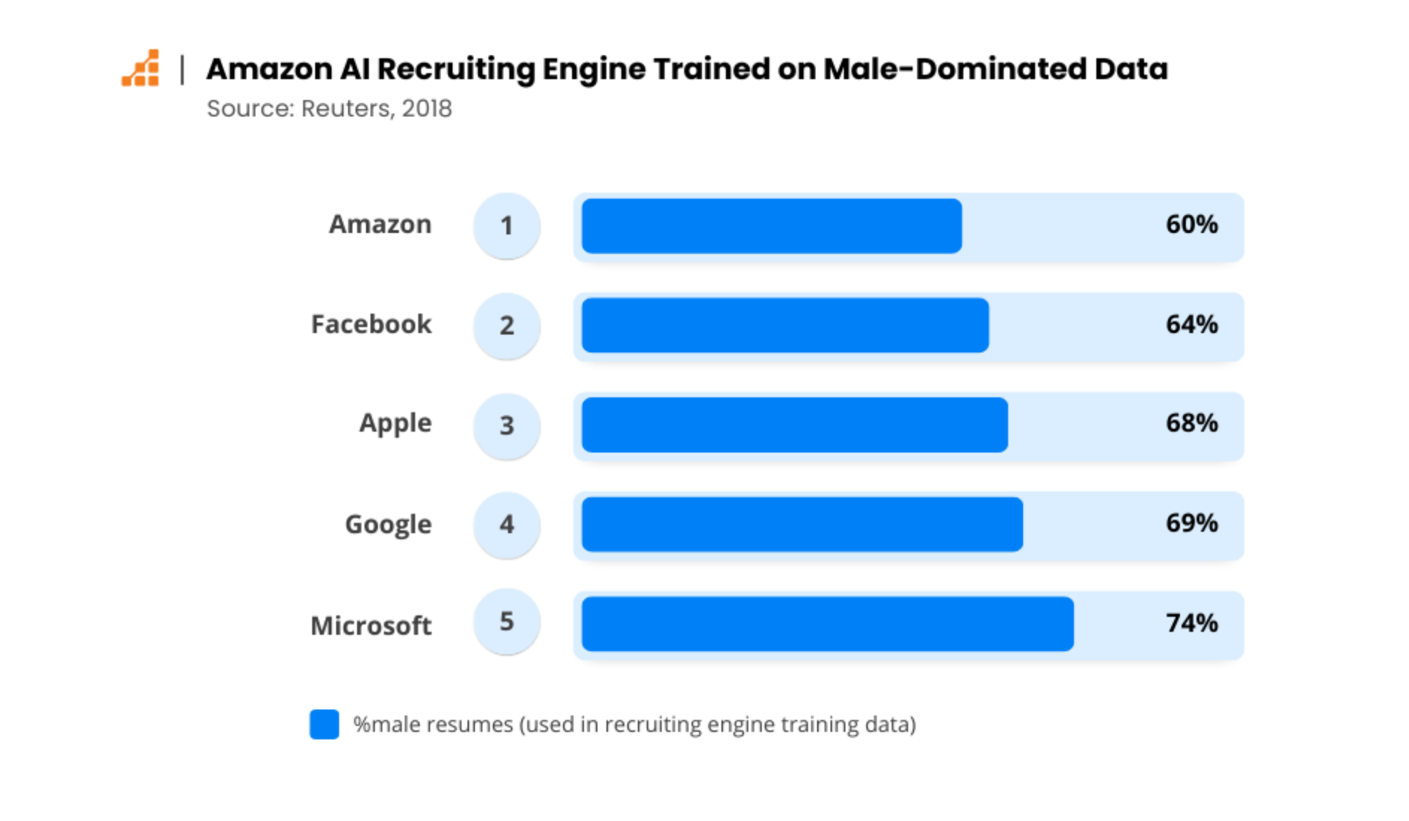
Amazon AI recruitment tool

- **Initiation and Objective:** Amazon began developing an AI recruitment tool in 2014 to streamline the evaluation of job applicants. The tool aimed to rate candidates on a scale from one to five stars, akin to product ratings on Amazon's platform. The goal was to create an efficient system that could automatically sift through hundreds of resumes and identify the top candidates quickly.

- **Training Data:** The AI system was trained using resumes submitted to Amazon over a ten-year period, predominantly from male candidates, reflecting the tech industry's gender imbalance. This historical data, steeped in existing gender biases, formed the foundation of the AI's learning process.

- **Detection of Bias:** By 2015, it became apparent that the system was not gender-neutral. The AI learned to penalize resumes containing terms like "women's," as in "women's chess club captain," and downgraded graduates from all-women colleges. It also showed a preference for resumes containing male-associated verbs such as "executed" and "capture".

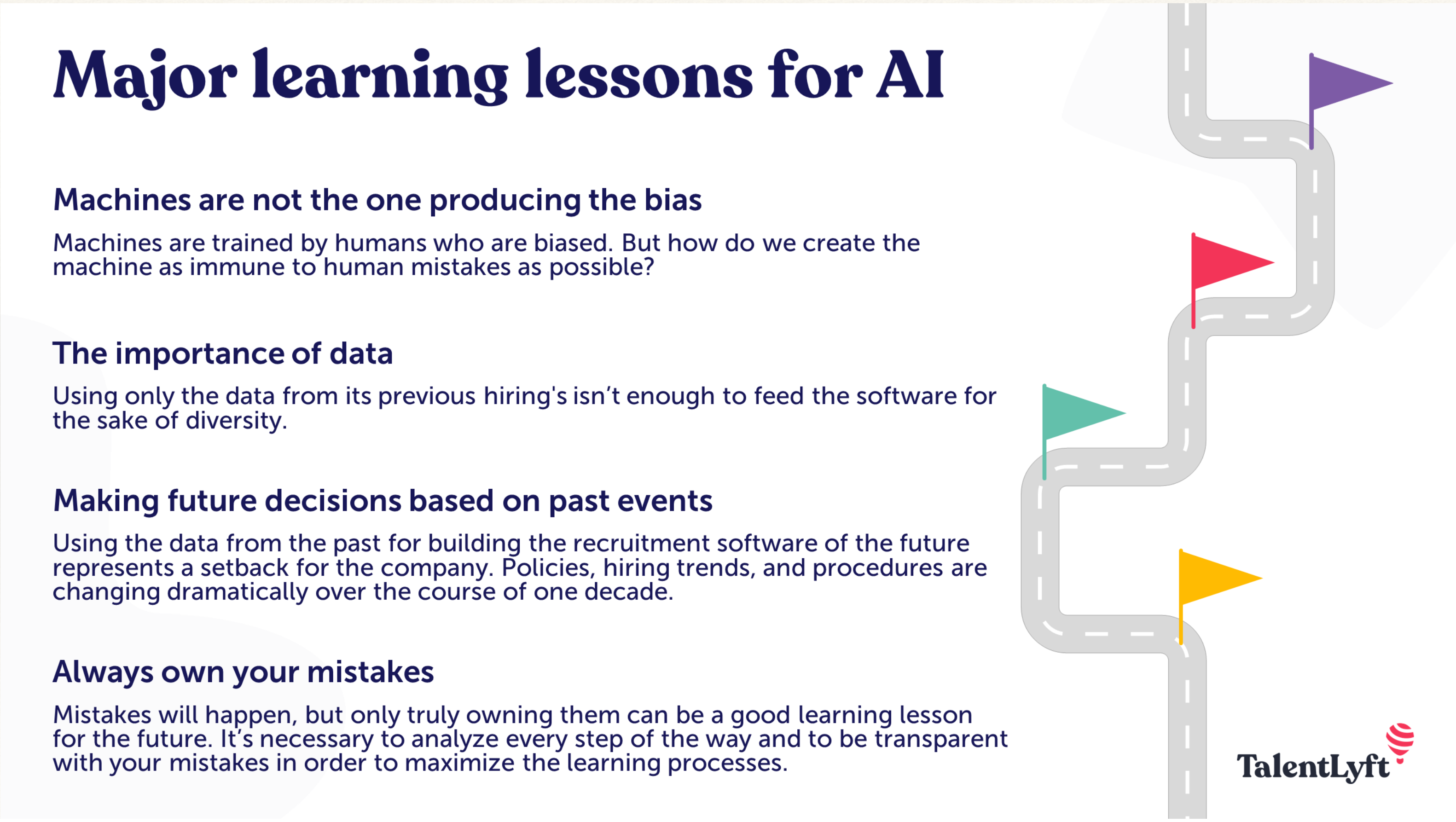
Amazon AI recruitment tool



Amazon AI Recruiting Engine Trained on Male-Dominated Data

Source: Reuters, 2018

Amazon AI recruitment tool



Major learning lessons for AI

Machines are not the one producing the bias

Machines are trained by humans who are biased. But how do we create the machine as immune to human mistakes as possible?

The importance of data

Using only the data from its previous hiring's isn’t enough to feed the software for the sake of diversity.

Making future decisions based on past events

Using the data from the past for building the recruitment software of the future represents a setback for the company. Policies, hiring trends, and procedures are changing dramatically over the course of one decade.

Always own your mistakes

Mistakes will happen, but only truly owning them can be a good learning lesson for the future. It’s necessary to analyze every step of the way and to be transparent with your mistakes in order to maximize the learning processes.

TalentLyft