

An abstract graphic on the left side of the slide, consisting of a network of thin, light-blue lines and small circles, resembling a circuit board or a neural network diagram. The lines are vertical and horizontal, with some diagonal connections, and the circles are small and evenly spaced along the lines.

COMBATING BIAS IN THE USE OF AI AND MACHINE LEARNING MODELS

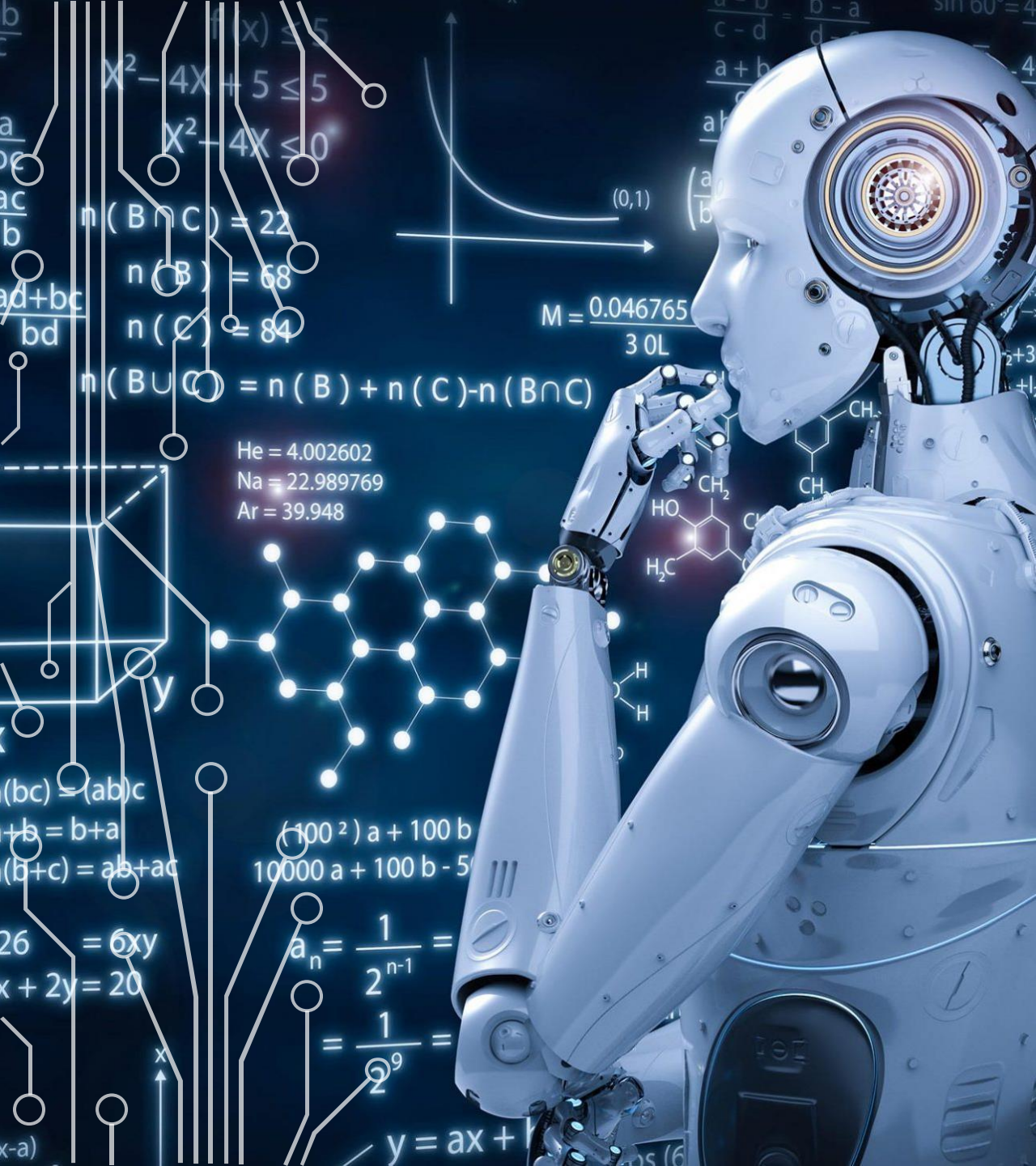
LIAN VAN OUDHEUSDEN

AGENDA

1. Understanding ethical bias in AI and machine learning models

2. Testing for bias

3. Additional considerations for promoting fairness



1. UNDERSTANDING ETHICAL BIAS IN AI AND MACHINE LEARNING MODELS

BIAS IN AI/ML HAS MADE HEADLINES...

Amazon AI Recruiting Tool

Amazon's AI-enabled recruiting software displayed bias against female applicants, having been trained on CVs of past successful applicants



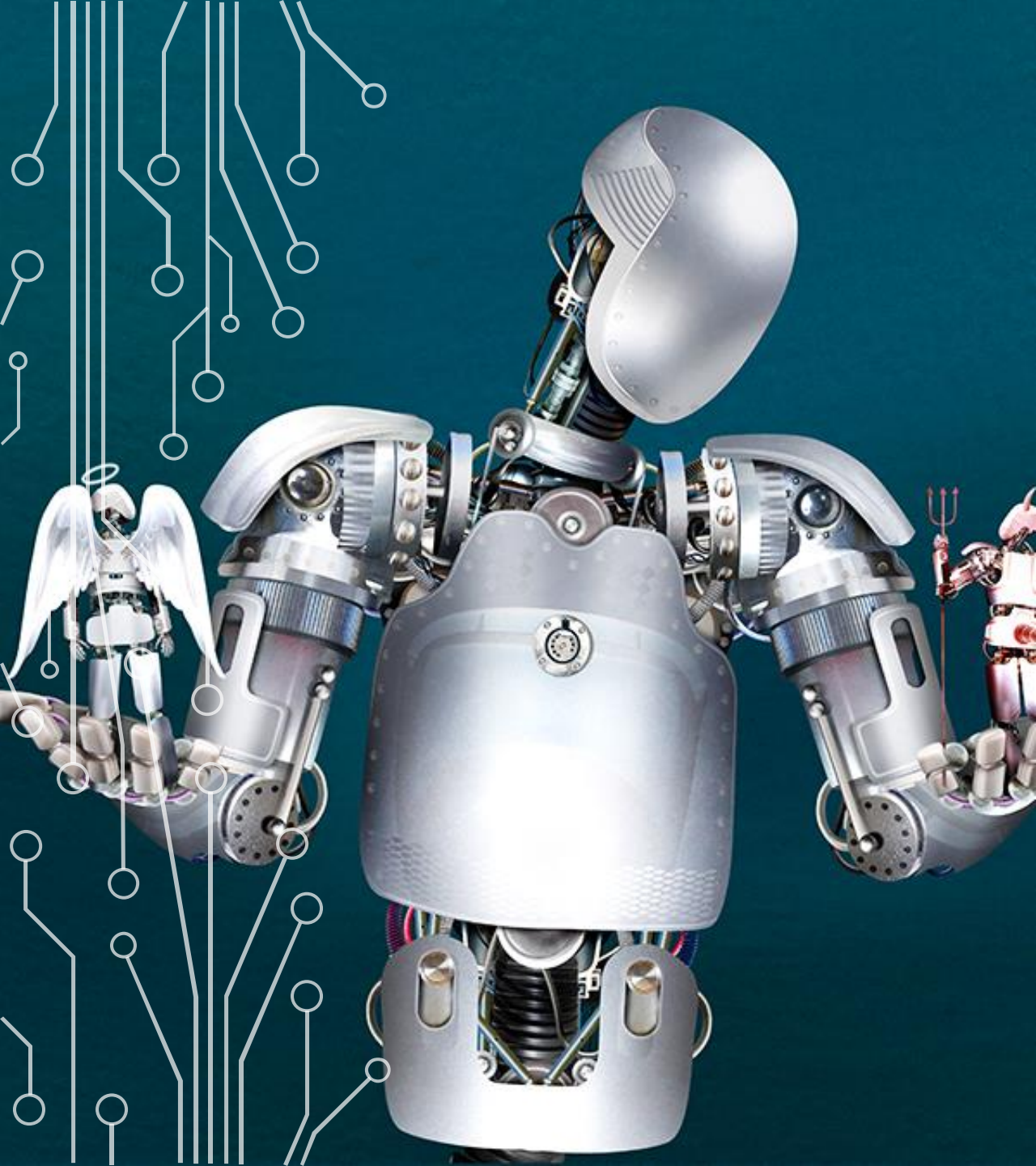
Researchers at the Georgia Institute of Technology found that Self-driving cars may be more likely to hit you if you have dark skin.

Self-driving cars

COMPAS

COMPAS was used in US court systems to predict the likelihood of recidivism and produced nearly twice as many false positives for black than white defendants.

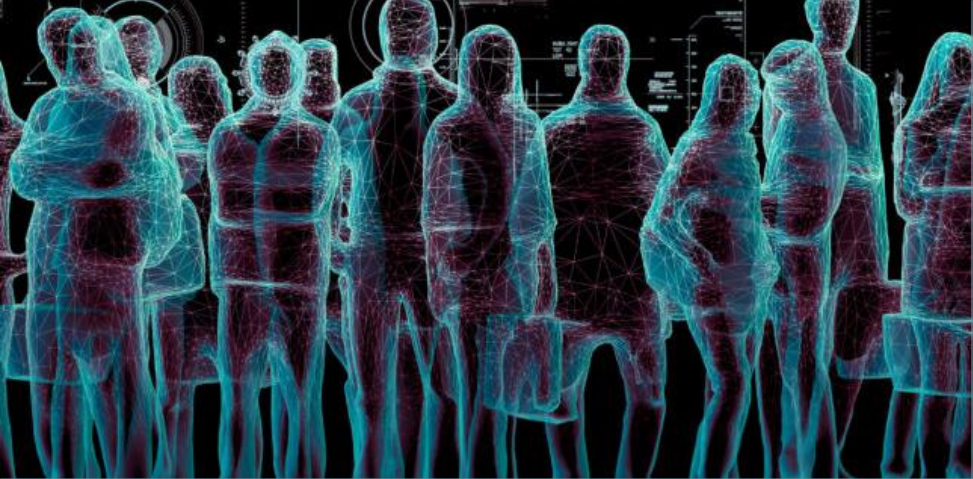




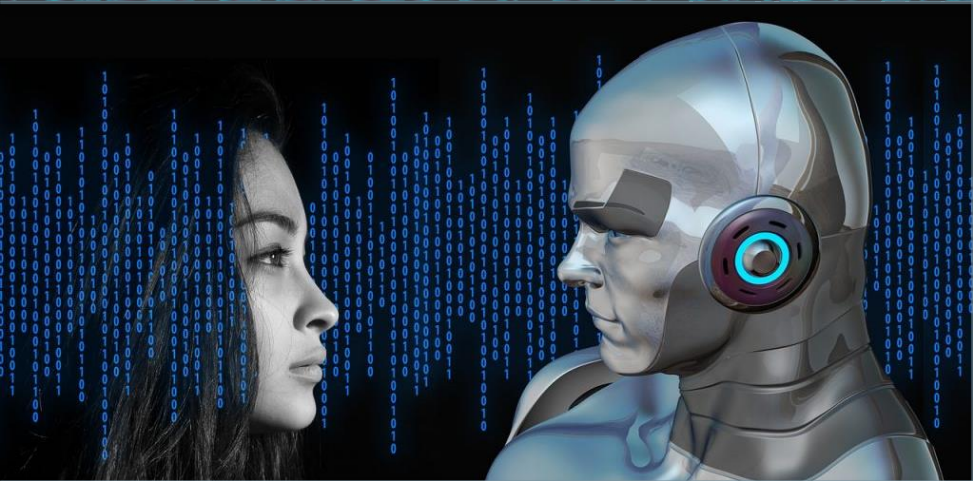
...BUT WHAT DO WE MEAN BY 'BIAS'?

“inclination or prejudice for or
against one person or group,
especially in a way considered to
be unfair.”

Oxford Dictionary



UNDERSTANDING SOURCES OF BIAS IN AI/ML IS KEY TO COMBATTING IT.



Selection bias:

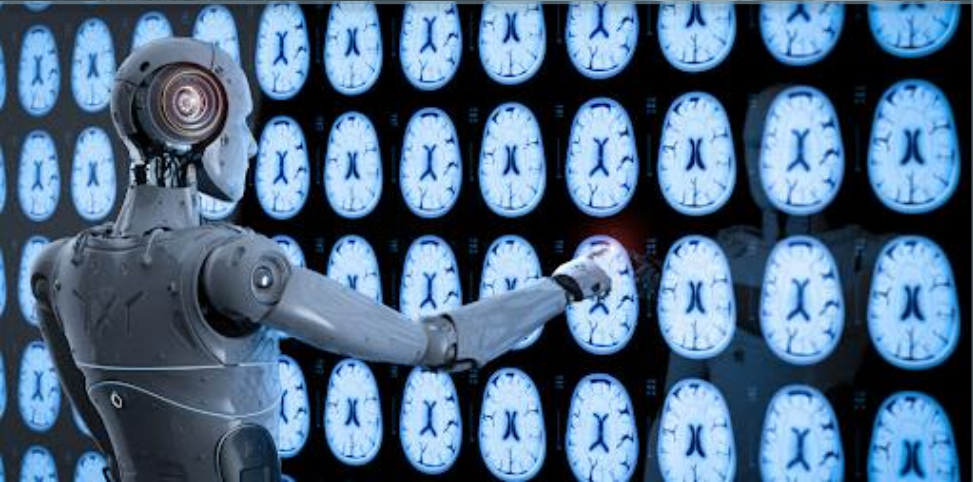
Results from data samples that are not representative of the total population

Historical bias:

Driven by biased decisions reflected in historical data

Group attribution bias:

Occurs when individual characteristics are extrapolated to entire groups



BANKING CARRIES POTENTIAL FOR BIAS IN COMMON AI/ML USE CASES...



- Credit decisioning



- Financial crime prevention



- Customer service

...AND FAILING TO ADDRESS THIS PROACTIVELY
LEADS TO LOSS OF TRUST.

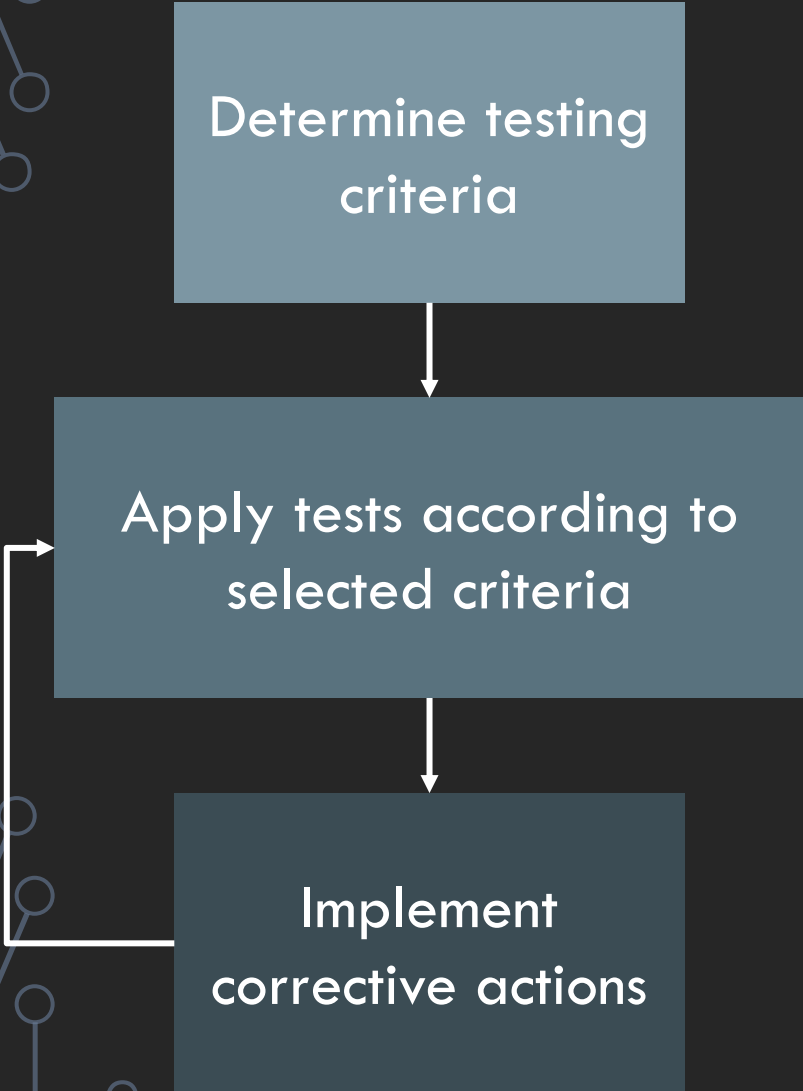


“The peculiar essence of our banking system is an unprecedented trust between man and man. And when that trust is much weakened by hidden causes, a small accident may greatly hurt it, and a great accident for a moment may almost destroy it.”

Walter Bagehot



2. TESTING FOR BIAS



```
graph TD; A[Determine testing criteria] --> B[Apply tests according to selected criteria]; B --> C[Implement corrective actions]; C --> B;
```

Determine testing
criteria

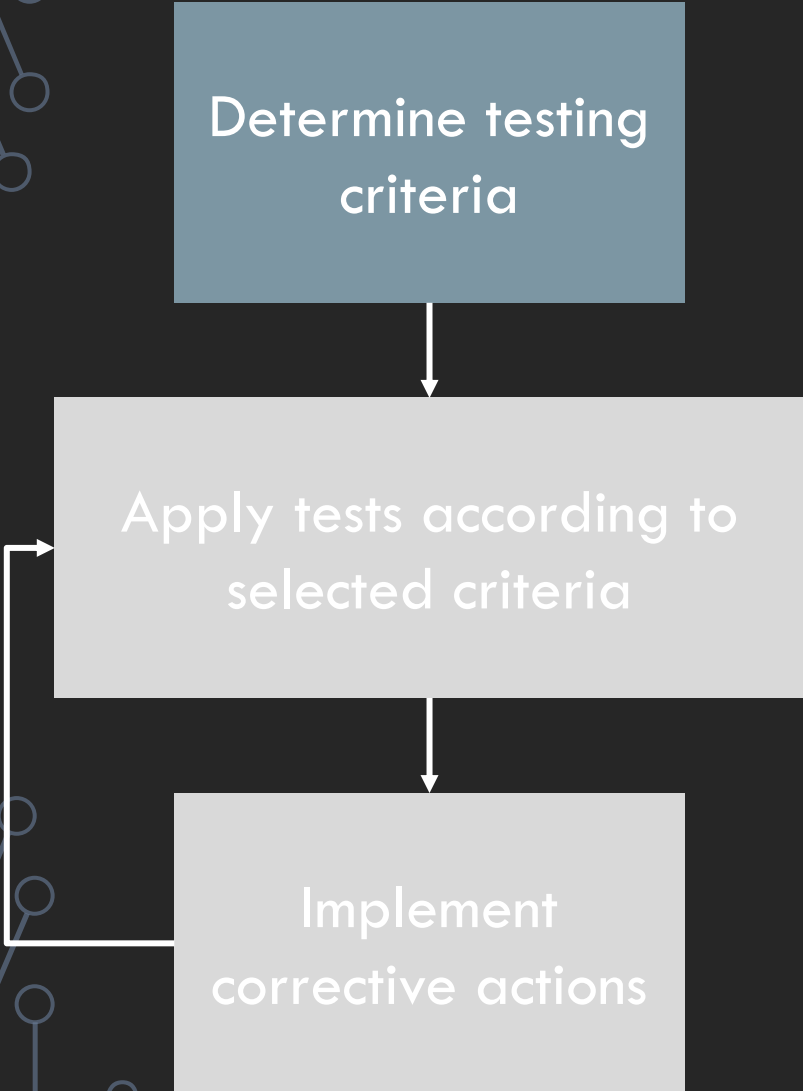
Apply tests according to
selected criteria

Implement
corrective actions

TESTING FOR AND ADDRESSING AI/ML BIAS SHOULD FOLLOW A SYSTEMATIC APPROACH.

Internal frameworks should articulate:

- The organisation's definition(s) of and stance on bias/fairness
- Criteria and thresholds for acceptable outcomes
- Tests to be applied for identification of bias
- Actions to be taken if test results not acceptable



```
graph TD; A[Determine testing criteria] --> B[Apply tests according to selected criteria]; B --> C[Implement corrective actions]; C --> B;
```

Determine testing
criteria

Apply tests according to
selected criteria

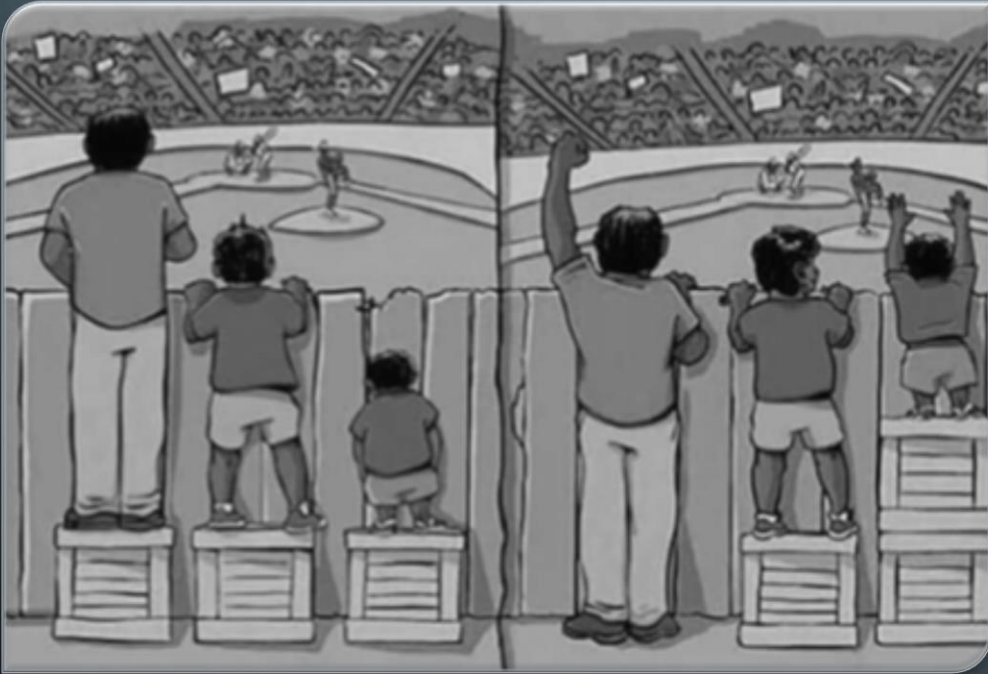
Implement
corrective actions

TESTING CRITERIA ARE ROOTED IN THE DEFINITION OF 'FAIRNESS'...

“impartial and just treatment or
behaviour without favouritism or
discrimination.”

Oxford Dictionary

... BUT VARIOUS DEFINITIONS OF 'FAIRNESS' CAN BE APPLIED IN THE CONTEXT OF AI/ML.



- **Individual fairness:** treating similar individuals similarly
- **Group fairness:** making the model's predictions or outcomes equitable across groups, particularly for potentially vulnerable groups



SELECTING A SUITABLE DEFINITION IS NOT TRIVIAL.

- Fairness criteria are often **mutually exclusive** – need to prioritise
- Selected definition needs to be translatable into **mathematical terms**
- Different **stakeholder perspectives** must be understood and considered
- Local **regulations** on direct and indirect discrimination must be adhered to

ORGANISATIONS SHOULD DEVELOP STANDARD TESTS FOR BIAS

Standard tests should:

- Align with selected fairness **definition(s)**
- Allow **comparison** between outcomes assessed against different definitions
- Be presented in a format that is **understandable and interpretable**, both internally and externally
- Include assessment against **defined thresholds**, where possible

```
graph TD; A[Determine testing criteria] --> B[Apply tests according to selected criteria]; B --> C[Implement corrective actions]; C --> B;
```

Determine testing criteria

Apply tests according to selected criteria

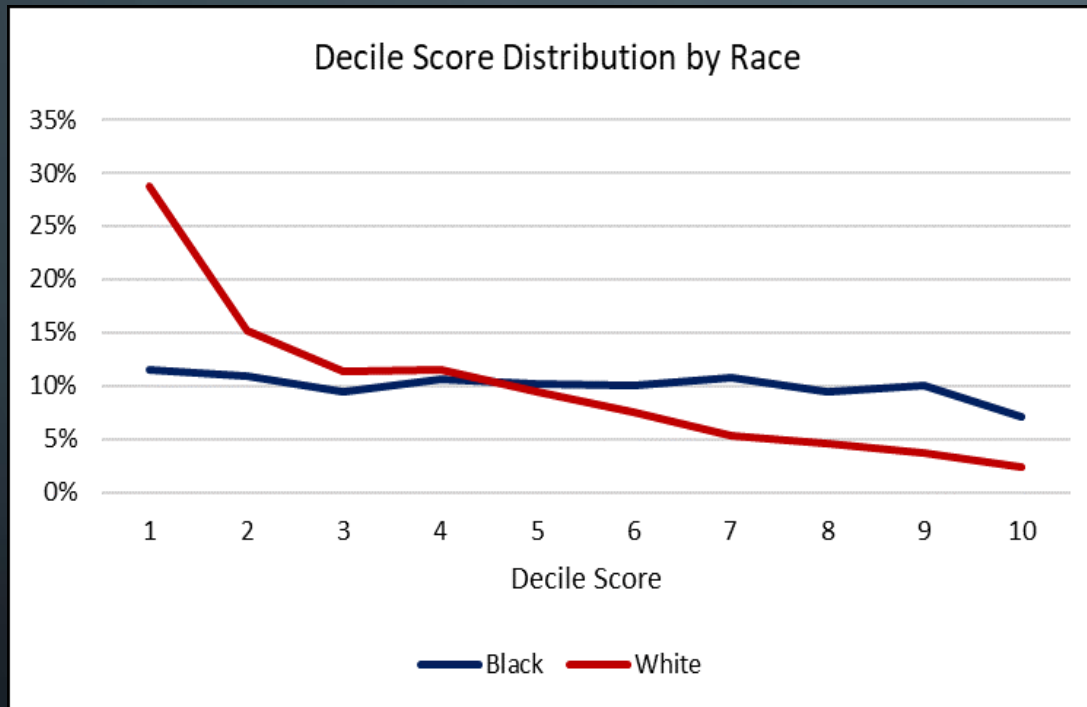
Implement corrective actions

USING FAIRNESS METRICS CALCULATED FOR THE COMPAS MODEL TO ILLUSTRATE:

Protected Variable: Race

Fairness Criterion	Description	Value	Rating
Equal Outcome (demographic parity)	Measures the extent to which the proportions of individuals from the protected and non-protected group that receive the desired outcome are the same	57.45%	MEDIUM
Equal Opportunity (equalised odds on true positives)	Measures the extent to which members of the protected class who actually deserve a desired outcome have the same chance of receiving it as those in the non-protected class	70.42%	LOW

CONSIDERING ADDITIONAL FAIRNESS METRICS:

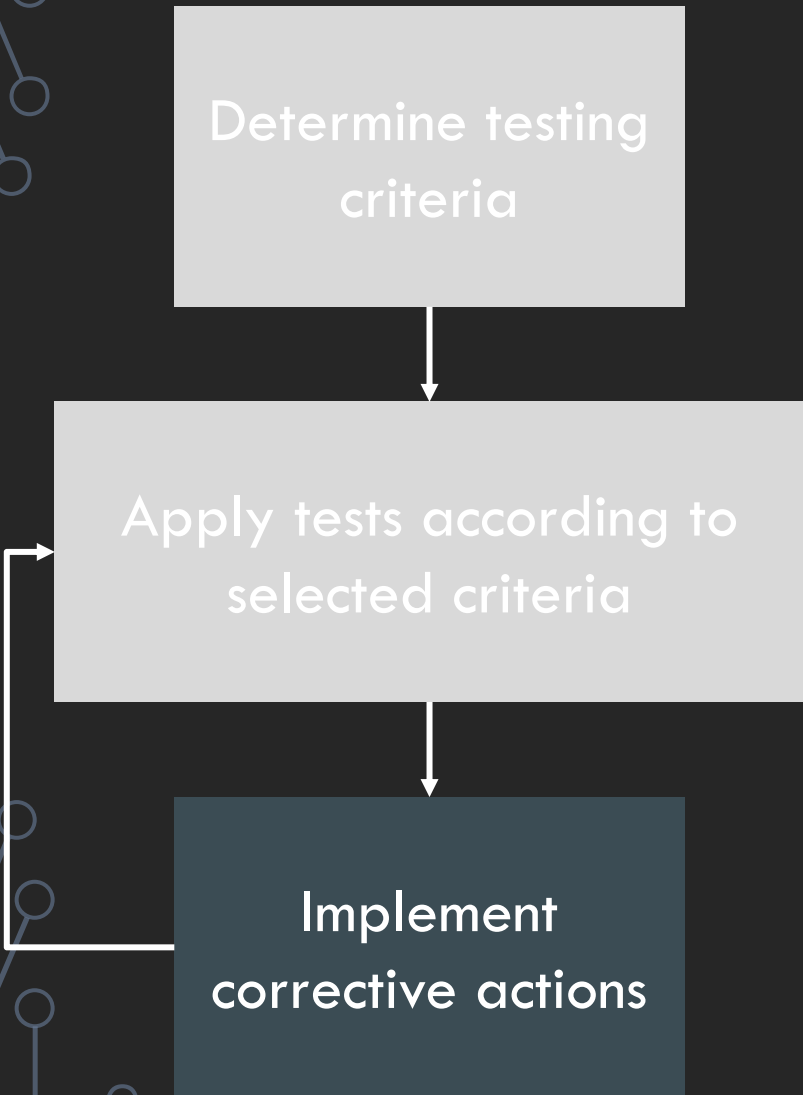


Fairness Metric	Value
False Positive Rate (FPR)	42.34% for Black and 22% for White

ORGANISATIONS SHOULD ALSO IDENTIFY APPROACHES FOR ADDRESSING BIAS ONCE IDENTIFIED.

Approaches to address bias should consider:

- Source of bias
- Fairness metric to be improved
- Acceptable post-intervention outcomes
- Potential impact on other fairness metrics



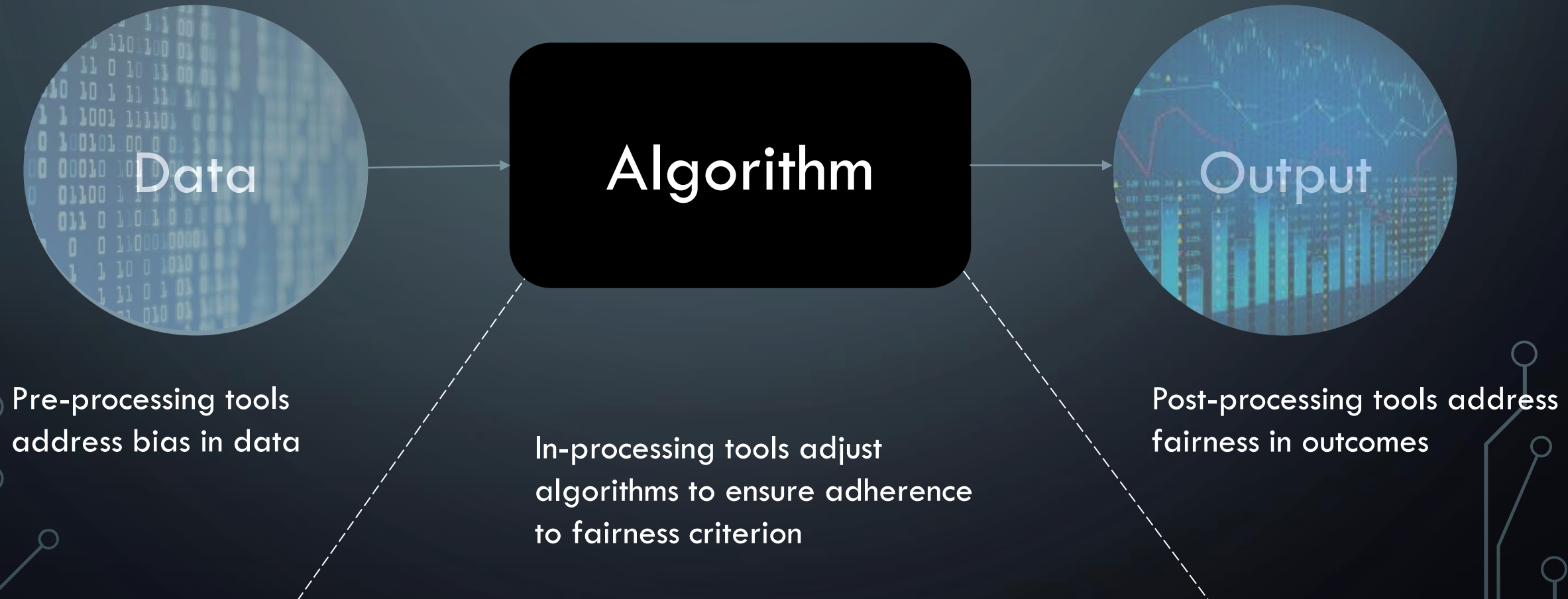
```
graph TD; A[Determine testing criteria] --> B[Apply tests according to selected criteria]; B --> C[Implement corrective actions]; C --> B;
```

Determine testing criteria

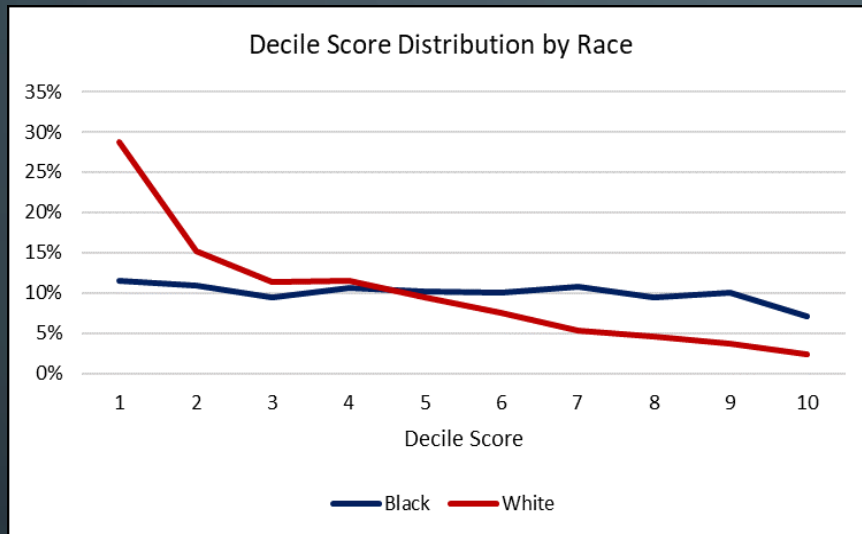
Apply tests according to selected criteria

Implement corrective actions

TOOLS AGAINST BIAS ARE TYPICALLY APPLIED
EITHER PRE-, IN- OR POST PROCESSING.



GOING BACK TO THE COMPAS EXAMPLE:



**Fairness
Metric**

Value

False Positive
Rate (FPR)

42.34% for Black and
22% for White

Solution	Intervention	Performance Metric	Value	Deterioration	Rating
Before	Equalised FPR	F1 Score	62%	-20%	Medium
After			50%		



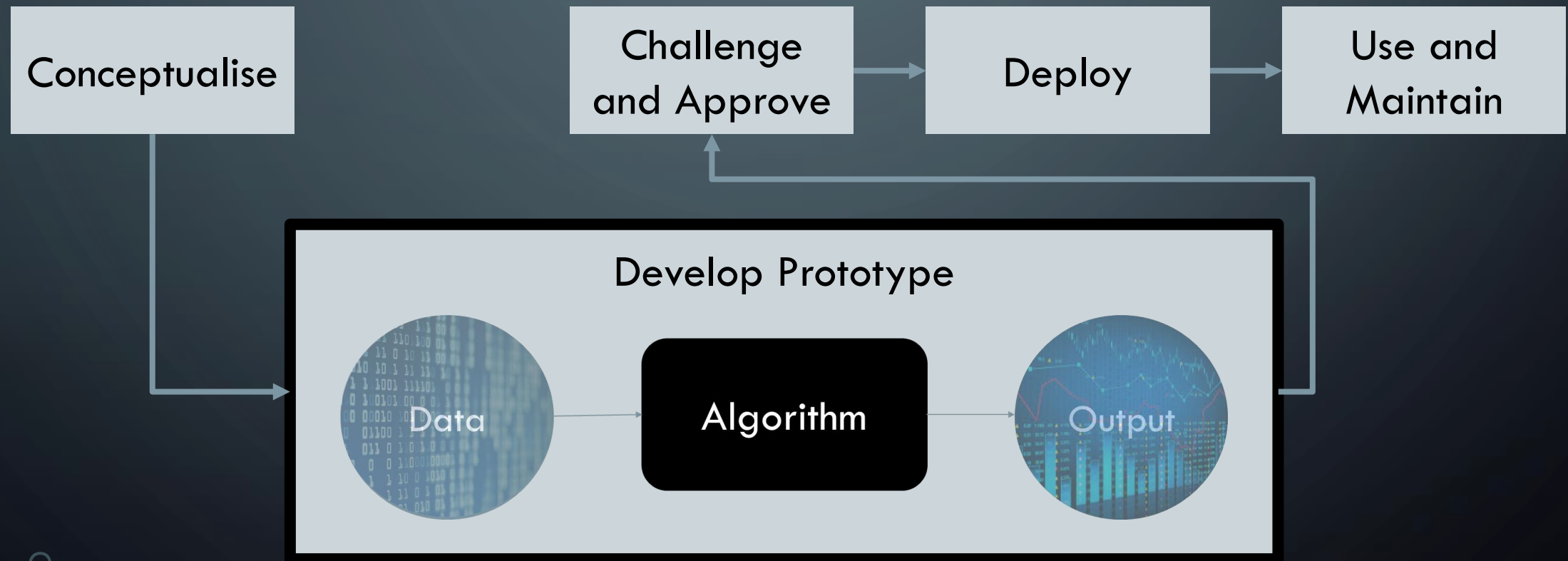
TOOLS TO ADDRESS BIAS WILL OFTEN LEAD TO LOSS OF ACCURACY.

Trade-offs between fairness and accuracy need to be carefully considered, taking into account the need for models to be **useful**, as well as the **real-world impact** of decisions based on models.



3. ADDITIONAL CONSIDERATIONS FOR PROMOTING FAIRNESS

TO PROPERLY PROMOTE FAIRNESS, WE NEED TO CONSIDER THE ENTIRE AI/ML LIFECYCLE...





HUMAN JUDGEMENT PLAYS A KEY ROLE IN BIAS AVOIDANCE...

Judgement is critical both for **supplementing** and for **interpreting technical tests** for bias.

Effective judgement needs to be supported by:

- Transparency
- Explainability
- Interpretability
- Accountability



... BUT THIS SHOULDN'T
REST ON THE
SHOULDERS OF ONE
MAN (OR WOMAN).



CLOSING REMARKS

“Rather an algorithm
that is monitored and
consistent in its bias,
than a human decision-
maker that is neither.”

Menelisi Maswanganye