AI and Data Fairness
Perspectives from Industry

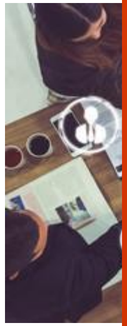
November 24, 2021

Dr. Sophia Ding, Senior Consultant

AI influences and determines our daily life

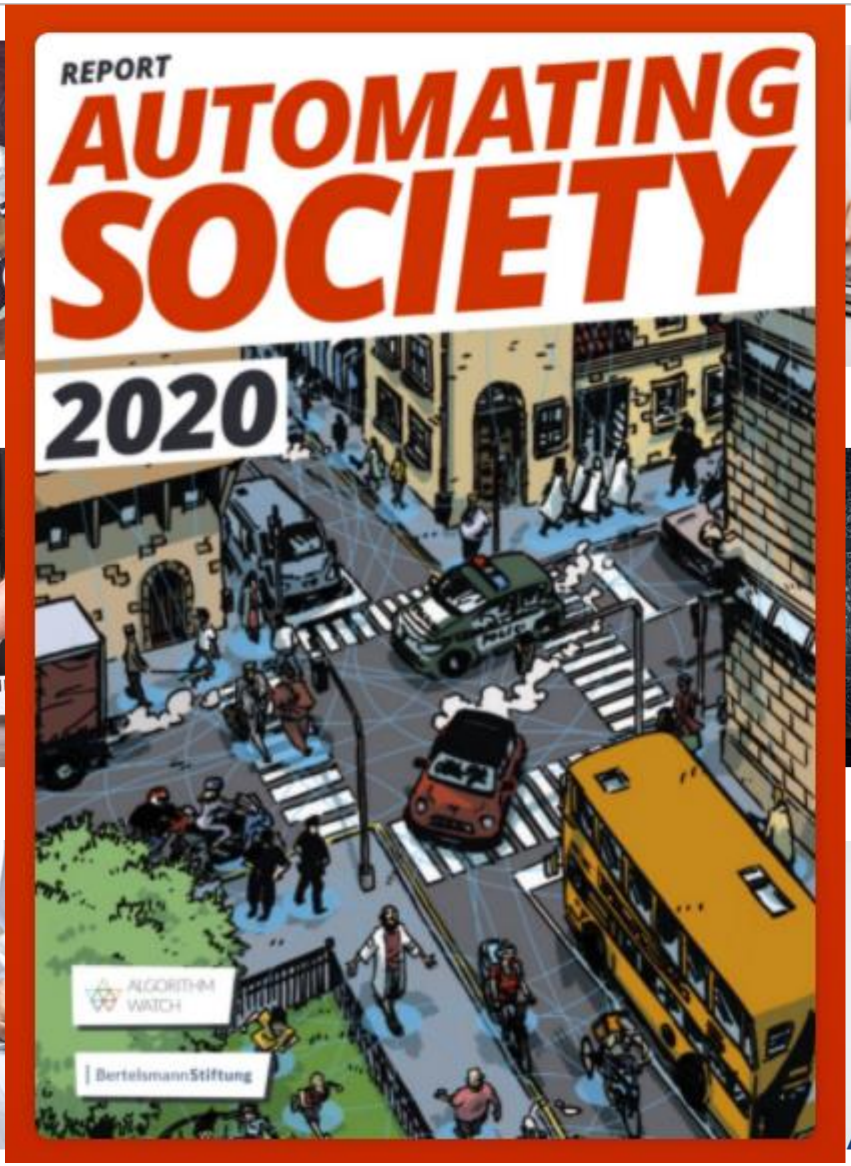
https://automatingsociety.algorithmwatch.org/

Human Resources:
- Application process (e.g. filmed interviews)
- pymetrics – candidate assessments (performance analytics)



Insurances:
- Risk assessments of clients
- Accelerated Claims Adjudication

Health Care:
- Length of stay at hospital
- Early detection of illnesses
- Treatment methods



Criminal justice:
- Probability of subsequent offence
- Probability of crimes while in prison

Education:
- Plagiarism checks (Turnitin software used by universities)
- Essay-grading Robo-readers

Marketing:
- Targeted or personalized ads for services and products (e.g. Amazon, Facebook)
- Political campaigning

The good…

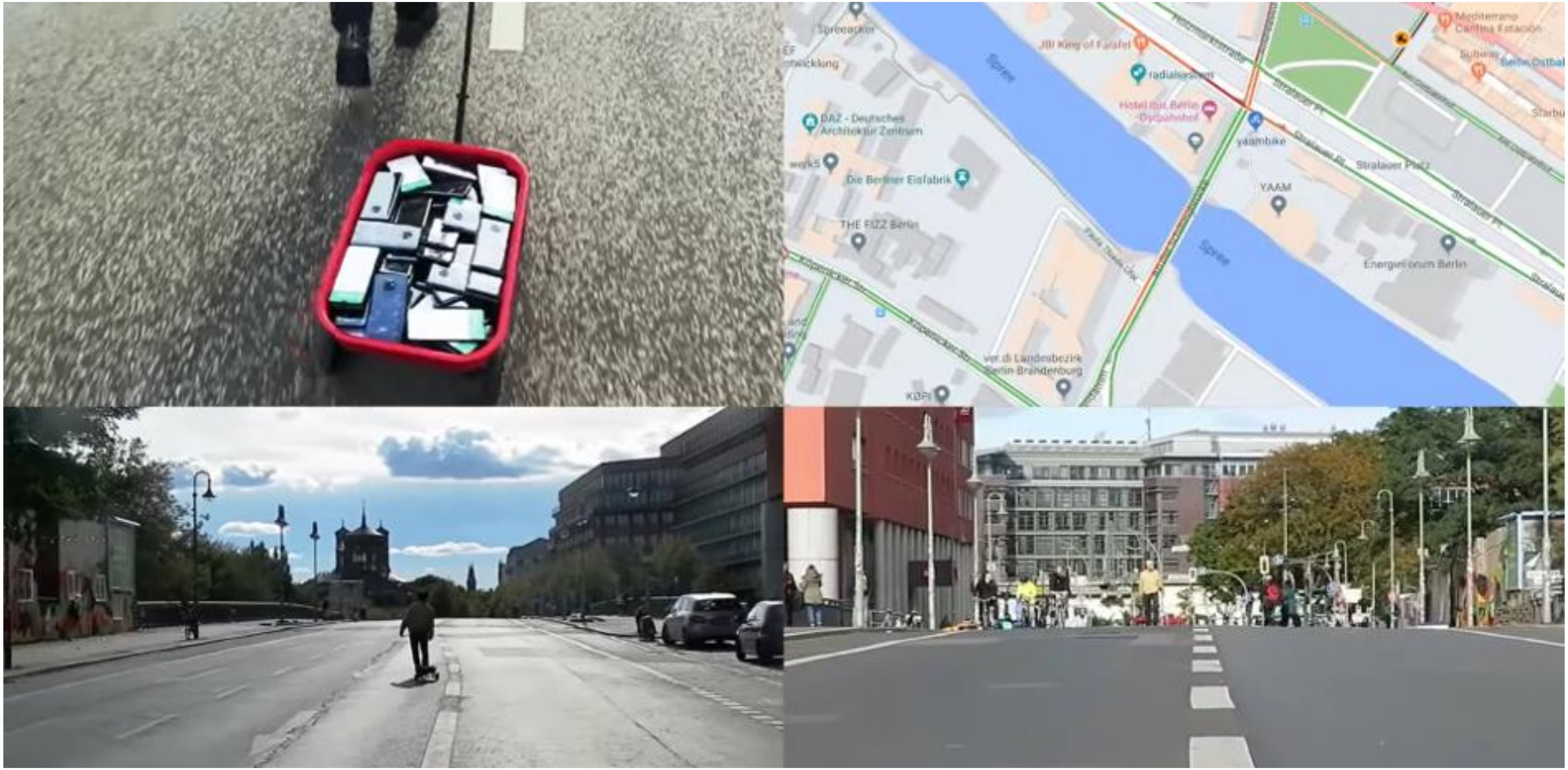
Faster and better drug discovery & better diagnostics

Improving the way we learn through personali-zation

Optimization of Energy Usage in Buildings

Optimizing location decisions regarding wildlife corridors

… the bad…


Spreeacker
DAZ - Deutsches Architektur Zentrum
Die Berliner Eisfabrik
THE FIZZ Berlin
Spree
Hotel ibis Berlin Ostbahnhof
yaambike
YAAM
Stralauer Platz
Energieforum Berlin
ver.di Landesbezirk Berlin-Brandenburg
KØPI
Subway

…and the ugly: AI used to determin high-school grades

Initial situation: cancellation of in-person A-level exams in the UK due to COVID-19 (March 2020).

Why AI?	Influencing Factors	AI Fail and negative impact	Audit perspective
• A-level grades are predicted by an algorithm • replacement of in-person exams	• School's historical grade distribution, student's rank within school, student's previous grades	• Disproportionately more lower grades for schools with a larger proportion of Black, Asian and Minority students • Public outcry • upward adjustment of marks for individual students according to teacher's predictions	• **However: the adjustment only fixes the symptoms** • **Decisive Audit-question: what is the under-lying cause?**

What is the plan for today?

(How) can we assess AI systems for fairness?

1. Qualitative Analysis
What does AI Fairness mean in the context of your use case?

2. Quantitative Analysis
(How) can we measure fairness?

- Intro to AI Trustworthiness
- AI Fairness in Practice
- Case Study

3. Risks mitigation
What are the risks for your organization and how can you mitigate them?

Intro to AI Trustworthiness

Terminology and common understanding: Artificial Intelligence and Machine Learning

Artificial Intelligence (AI)

software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with

As defined by the EU AI Act – regulation proposal

Logic- and knowledge-based approaches, e.g. expert systems

Statistical Approaches, Bayesian Estimation, Search and Optimization methods

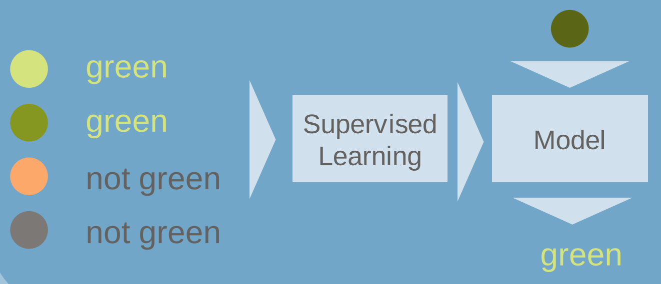
Machine Learning (ML)

Algorithms that enable an artificial system to learn from experience and to generalize it.

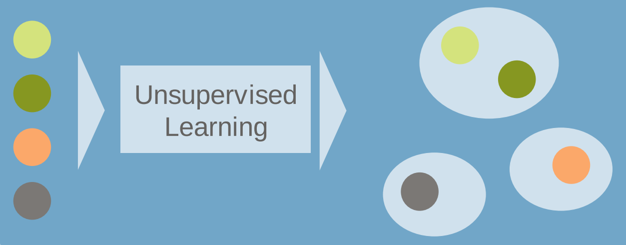
Reinforcement Learning

Deep Learning

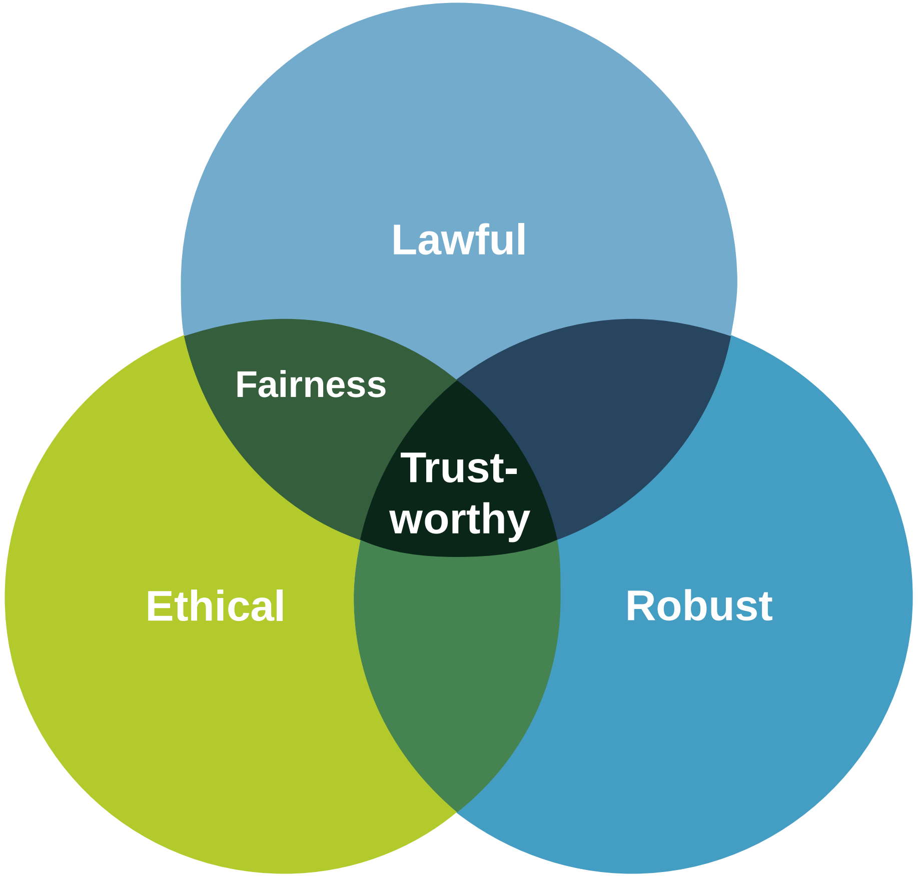
Supervised Learning



Unsupervised Learning



Mitigating AI risk: Trustworthy AI



Trustworthy

- AI services should be ***trustworthy*** - throughout the entire life cycle

Lawful

- AI services should be ***lawful*** and comply with all applicable laws and regulations

Ethical

- AI services should be ***ethical*** and ensure compliance with ethical principles and values

Robust

- AI services should be ***robust***, both from a technical and social point of view, as AI systems can cause unintended harm even with good intentions

Source: Ethics Guidelines for Trustworthy AI – High-level Expert Group on Artificial Intelligence

Secure

- AI services should be ***secure***, on the one hand against technical attacks on the availability and against manipulation of the algorithm

AI is not the solution for all of our problems, but if we decide it makes sense to use this technology, we should make sure it is trustworthy.

Three phases of governing AI

1	2	3
General principles for AI (Soft law / self regulation)	Technical tool boxes to implement AI principles	Regulation

1 — General principles for AI (Soft law / self regulation)

- Ethics Guidelines for Trustworthy AI
- OECD AI Principles
- Principles written by companies
- … and more than 150 more

2 — Technical tool boxes to implement AI principles

- IBM AI 360 Fairness
- Aequitas
- AWS SageMaker Clarify
- Microsoft Fairlearn

3 — Regulation

- EU AI Act

A Framework for AI Risks – Fairness is only one aspect that contributes to trustworthiness

Governance for the use of decision-making algorithms

Industry Specific Factors				
	G01 company's risk appetite	G02 leadership engagement and steering	G03 management and reporting structures	G04 compliance and governance
	G05 data protection principles (by privacy or by design)	G06 general and specific policies and directives	G07 documentation and traceability	G08 courses and awareness measures

AI/ML-specific risk areas

R01 Fairness and transparency	R02 process accuracy	R03 security protection for the application	R04 minimization and appropriation of personal data	R05 transparency of the result's development
R06 social discrimination	R07 loss of accountability	R08 manipulation and malicious use	R09 complexity-related control loss	R10 use of counter or attack algorithms

AI Fairness in Practice

There are different sources for bias in AI systems (this list is not exhaustive)

Implicit	By Design	Temporal	Qualified

A Short Introduction into AI Fairness – How to Deal with It

Part 1: Qualitative Analysis

"What does Fairness mean in the context of the use case?"

1

Understanding the algorithm and its application area
"What does the algorithm do?"

2

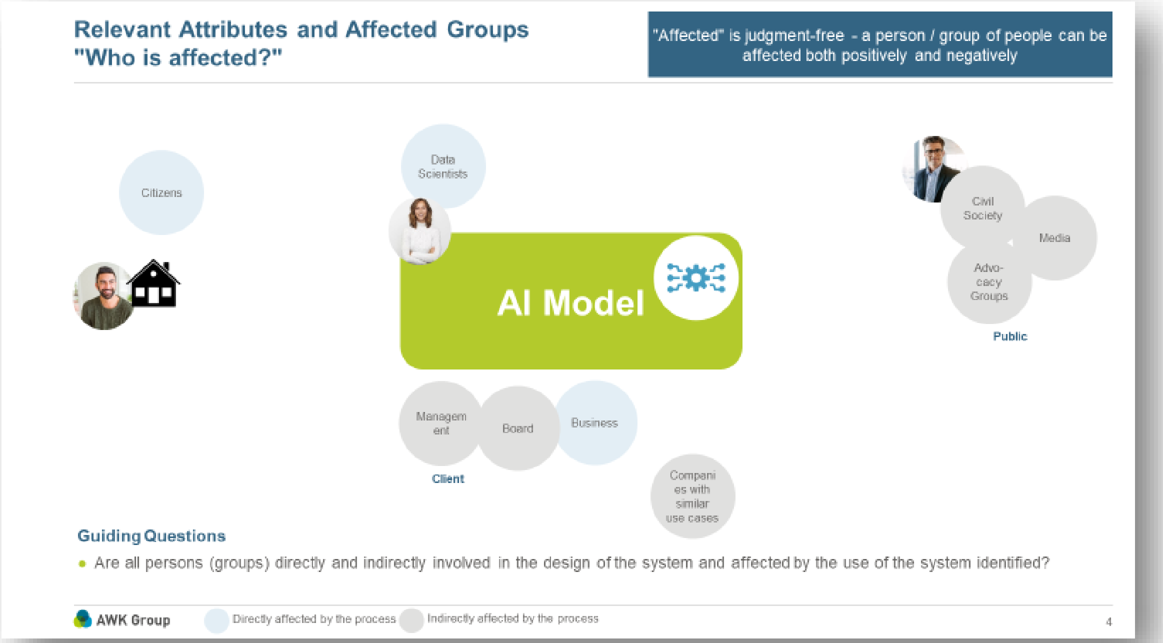
Relevant attributes and affected groups
"Who is affected?"

3

Estimating the consequences of unwanted bias
«How bad would damage be?»

4

Risk assessment
"How probable is damage?"



A Short Introduction into AI fairness – How to Deal with It

Part 2: Quantitative Analysis

"Is there statistical evidence for bias?"

5

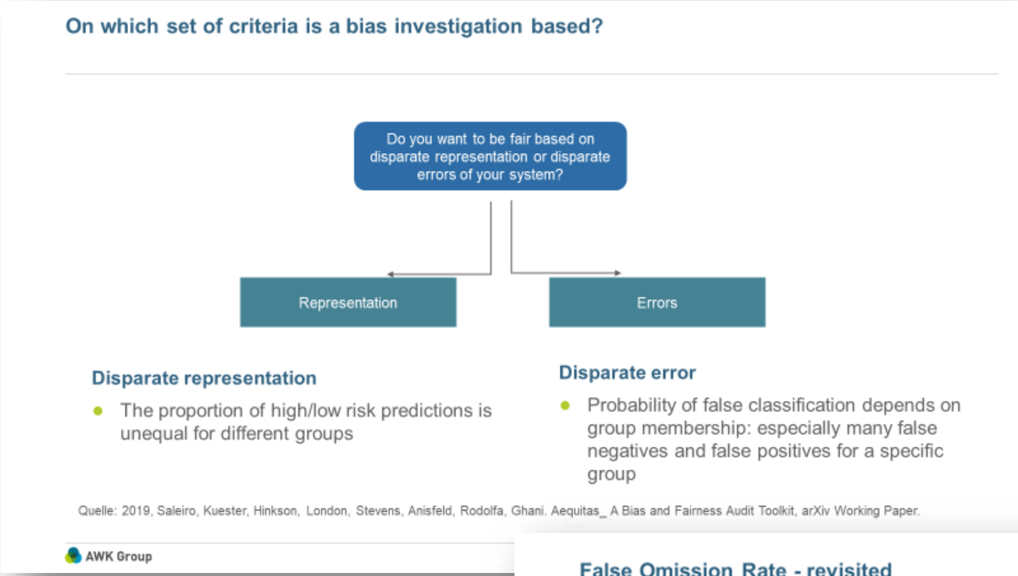
Selection of the relevant bias metrics

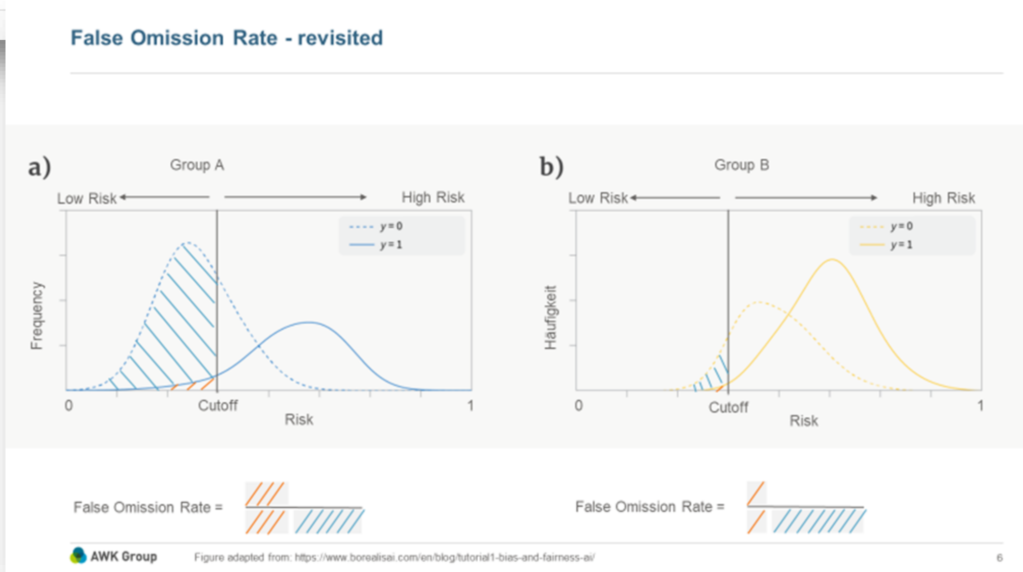
"Which metric is meaningful for the use case?"

6

Interpretation of key figures given by the qualitative analysis

"For which groups of people are there signs for a wanted or unwanted bias?"



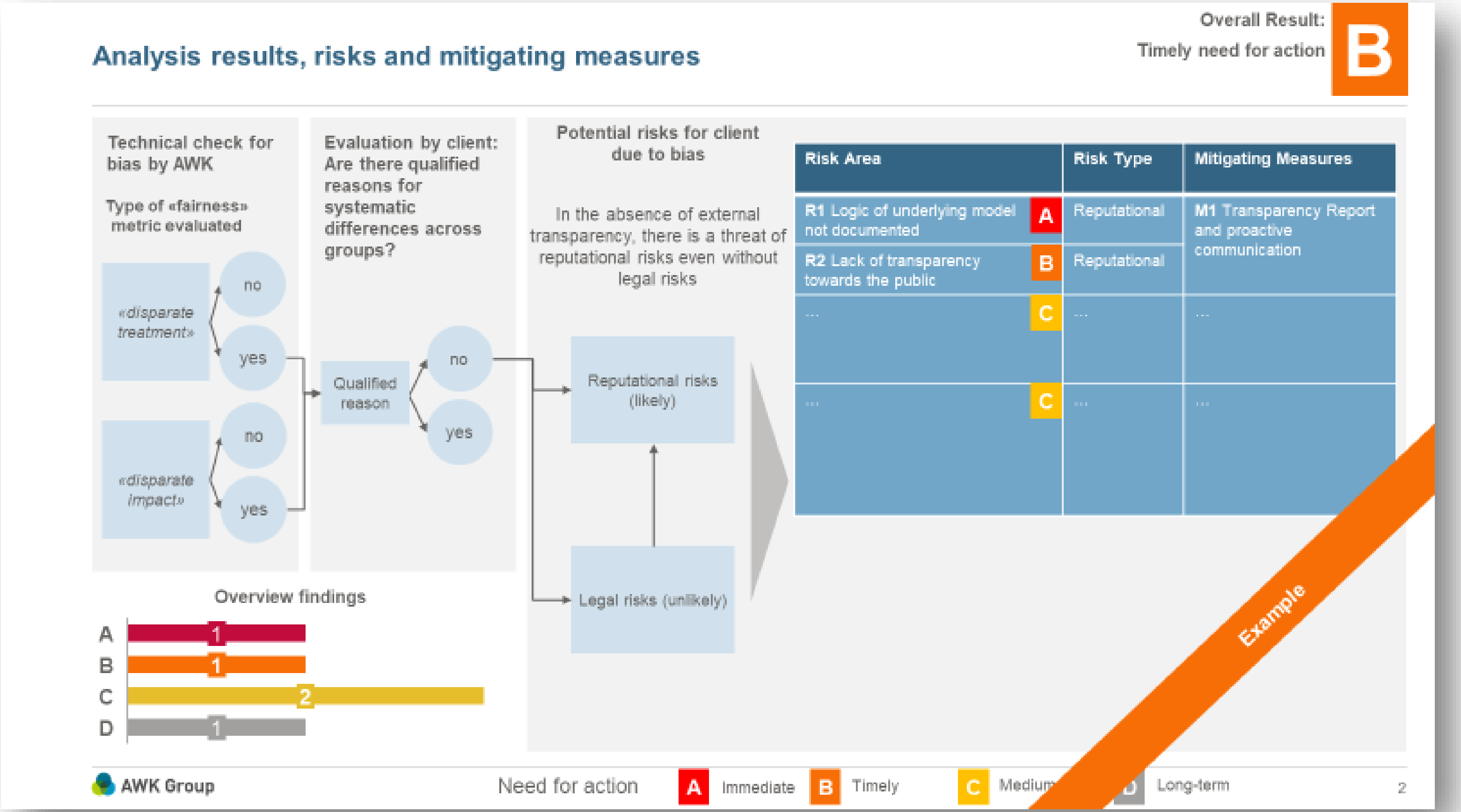


A Short Introduction into AI fairness – How to Deal with It

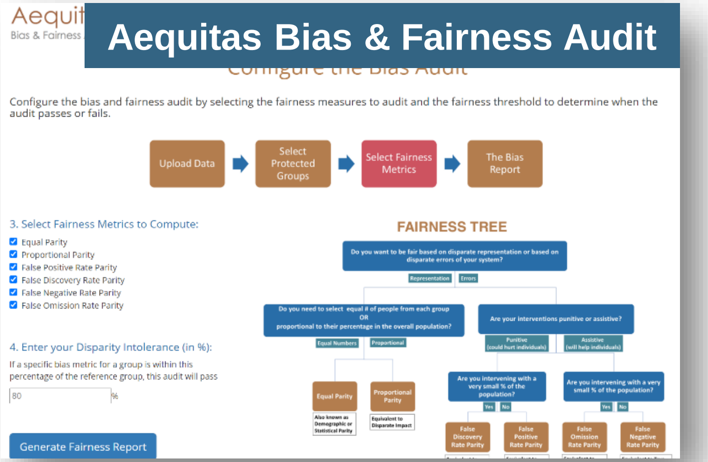
Part 3: Deriving risks and measures to mitigate risks

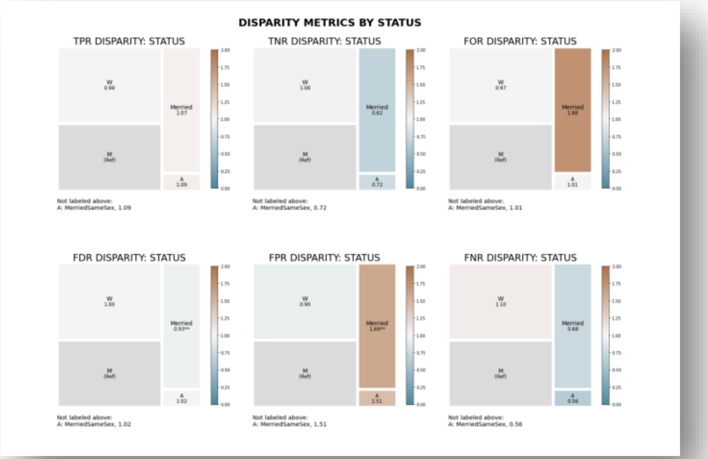
7

Identifying risks and counter-measures
“Which risks and which mitigating measures exist?”



We use the Aequitas Toolbox for a quantitative risk assessment





- Open source toolbox from the academic environment (Data Science for Social Good Project, University of Chicago)
- Allows evaluation of different fairness metrics
- Plus point: Was specifically designed for audits and contains a use case-oriented decision tree with regard to the selection of the appropriate fairness metrics as well as detailed explanations of the audit results

http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

Today

- We will perform the quantitative assessment on a real world use case

Case Study

Use Case: COMPAS

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

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

More Terminology

Question: What is the problem with false negatives? What is the problem with false positives?

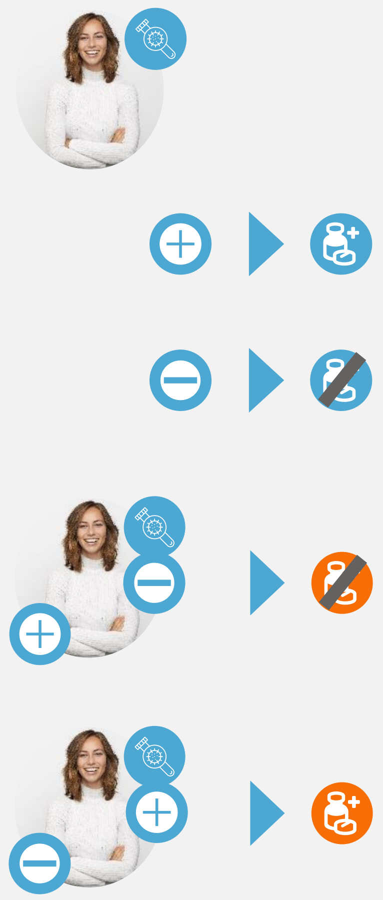
COVID-19 Testing

Sick → Intervention: Treatment

Not sick → No Treatment

Sick → No Treatment = False negative

Not sick → Treatment
False positive



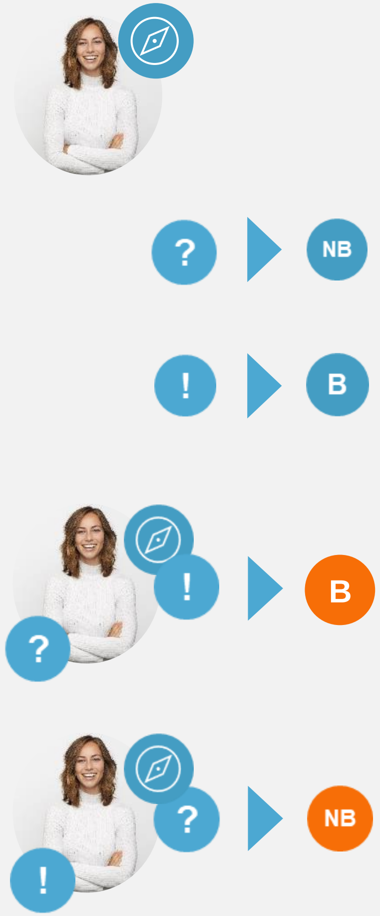
COMPAS

High Risk for Recidivism → Intervention: Cannot be bailed out

Low Risk → can be bailed out

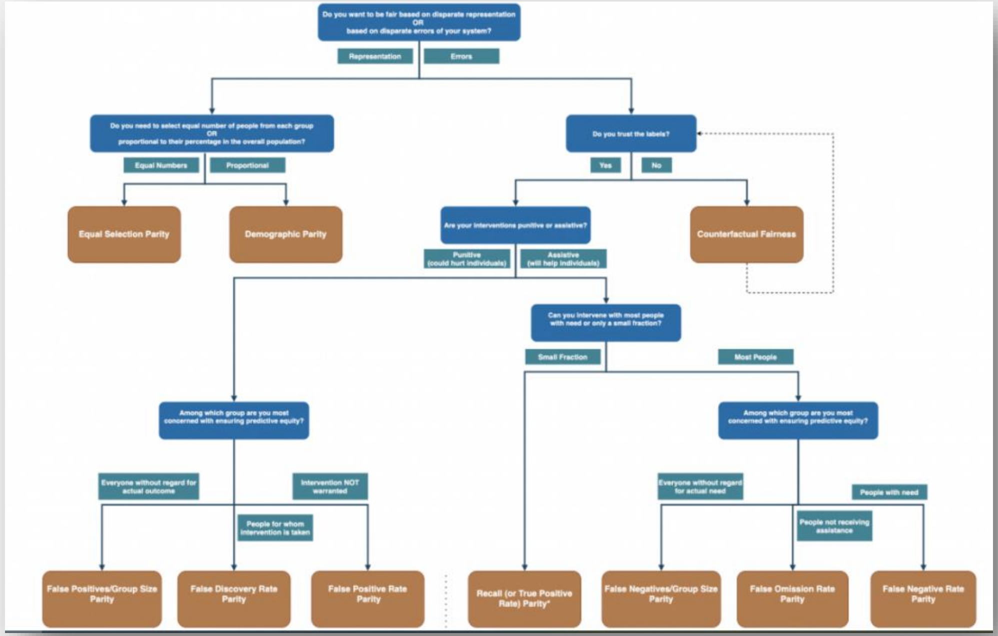
High Risk → can be bailed out = False negative

Low Risk → cannot be bailed out
False positive



? High risk for recidivism ! Low risk for recidivism NB Cannot be bailed out B Can be bailed out

What are fairness metrics and how do they differ?



Aequitas decision tree:
Which metrics are relevant for COMPAS?

Questions characterizing the use case and the understanding of fairness

Possible responses

Fairness Metrics

Based on the decision tree, we can determine which fairness metrics make sense in our use case and should be evaluated in more detail.

There are two types of fairness metrics

Equal representation

«Only predicted risks for recidivism matter»

Equal errors

«Both the predicted risk for recidivism and the acutal risk for recidivism matter»

Equal representation: Details

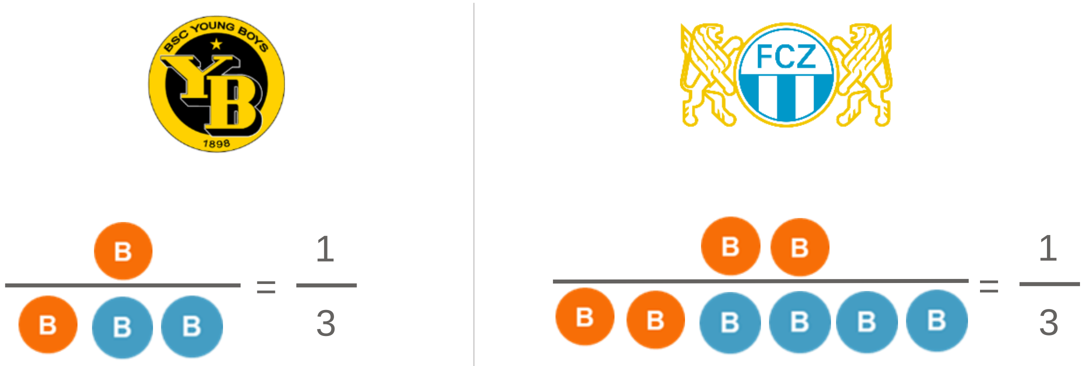
Fairness as **equal representation**:

- The actual risk does not matter.
- Fair outcome: the same share of YB-Fans can get out on bail as of FCZ-Fans.
- Underlying Assumption: An unequal distribution of risks across groups is based on existing social biases.
- We call these metrics «bias transforming»

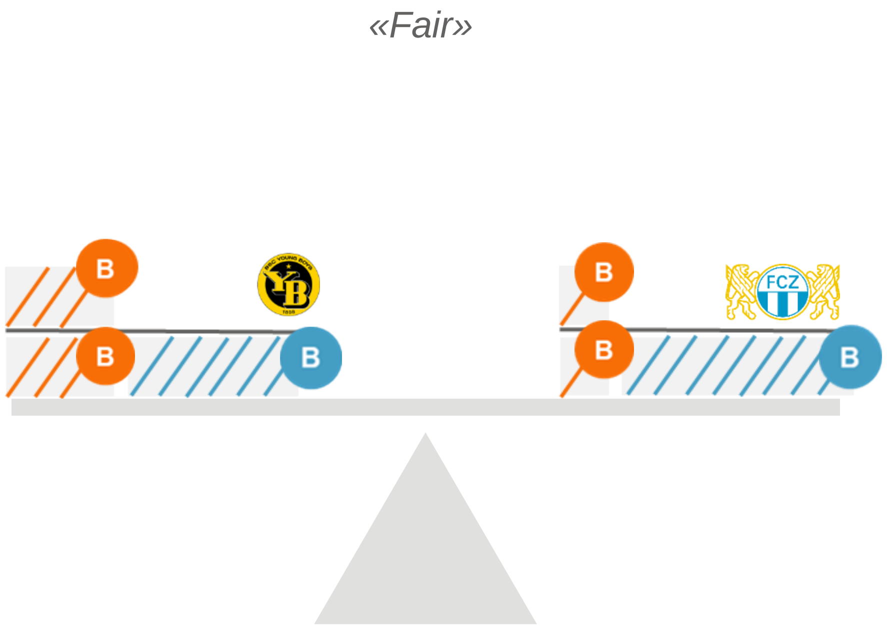
Equal errors: Details

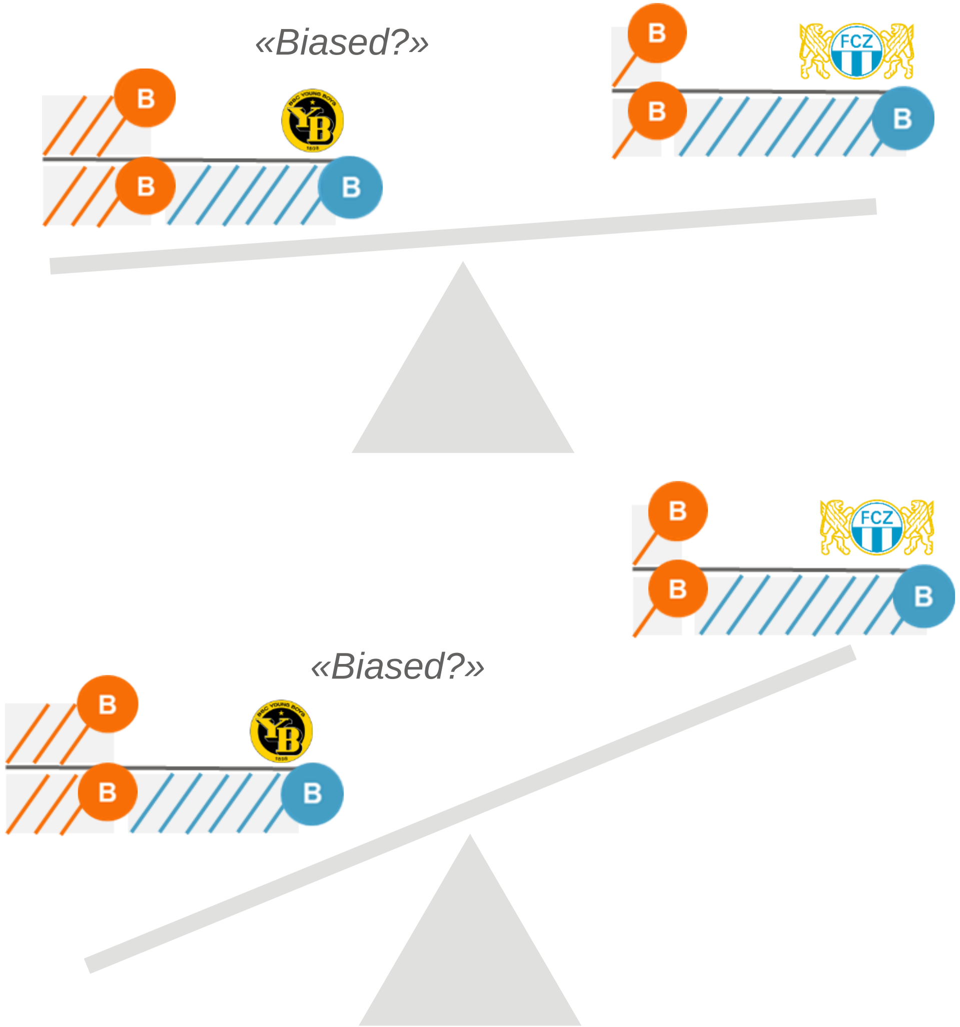
Fairness as **equal errors**:

- Only the false positives / false negatives matter.
- Fair outcome: The probability of a false negative / false positive is the same among YB-fand and FCZ fans. hoch ist.»
- There are several metrics which differ slightly, e.g. false omission rate (probability that I am a high risk for recidivism and can get out on bail.



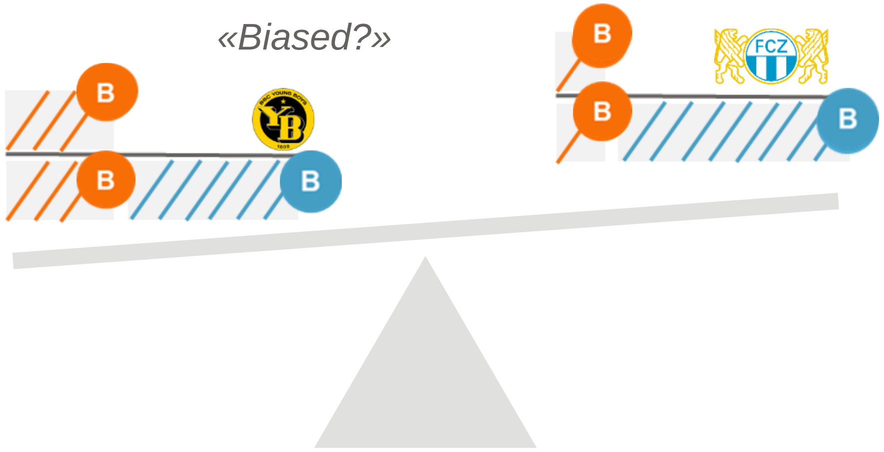
Measuring fairness: Some challenges

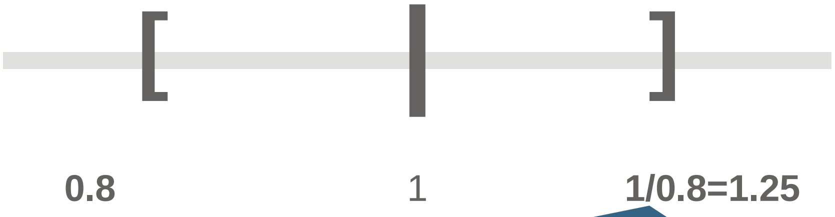




In order to assess unbiasedness, we need to determine at what point two values are close enough to each other to be considered unbiased.

Measuring fairness: Some challenges

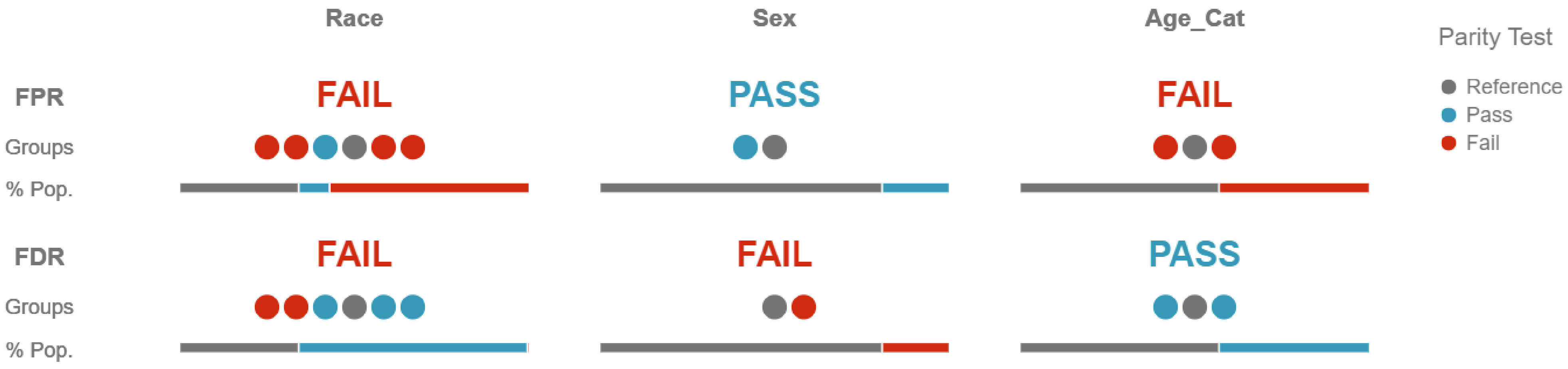




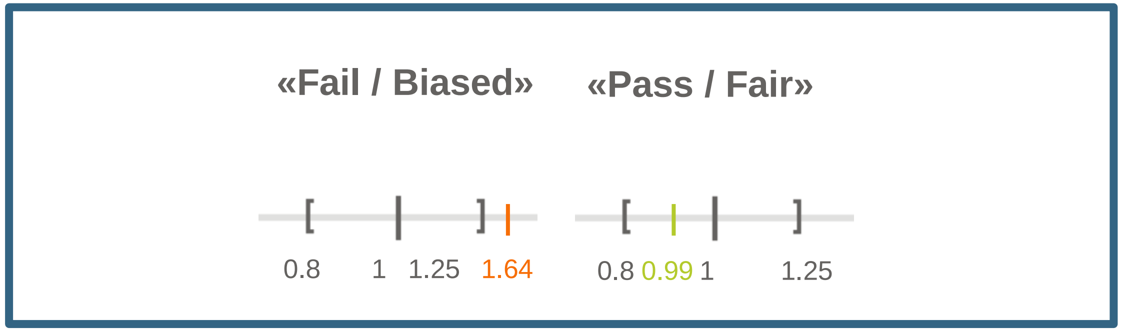
80% rule, i.e. the proportion of defendants who can get on bail in one group may not be lower than 80% of this proportion in the other group.

Source 80%-Rule: 1978 Uniform Guidelines on Employee Selection Procedures, U.S. Equal Employment Opportunity Commission (EEOC)

Measuring fairness: Some challenges



For a group to pass the parity test its disparity to the reference group cannot exceed the fairness threshold (1.25).
An attribute passes the parity test for a given metric if all its groups pass the test.



Measuring fairness: Some challenges

Protected Attributes

- Features that are not allowed to be used as the basis for decision-making.
- Either given by law (e.g. anti-discrimination law) or because of a company’s or institution’s values (e.g. code of ethics).
- Examples: gender, race, religion, gender, marital status, age, nationality, and socioeconomic status.

Reference Group

- Group which is taken as the baseline.
- Often this is the group that was historically advantaged in the context of the use case.
- Other options: majority group (largest group), lowest value of the fairness metric

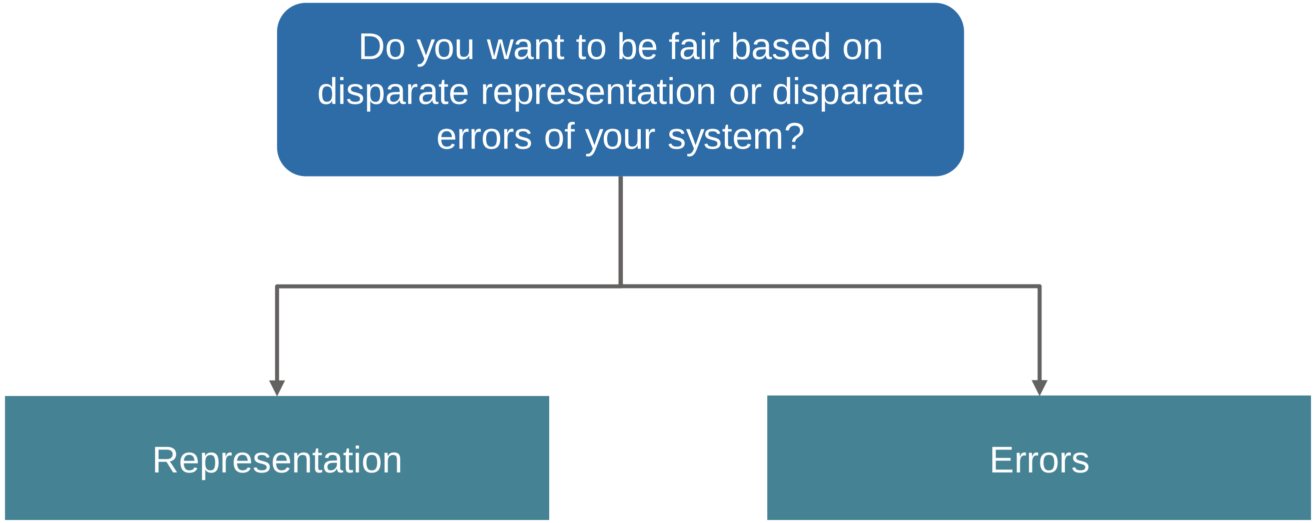
Who makes these choices?

- Technical choices embody ethical values.
- Stakeholders should be included in the qualitative analysis that leads the quantitative analysis.

Use Case: COMPAS

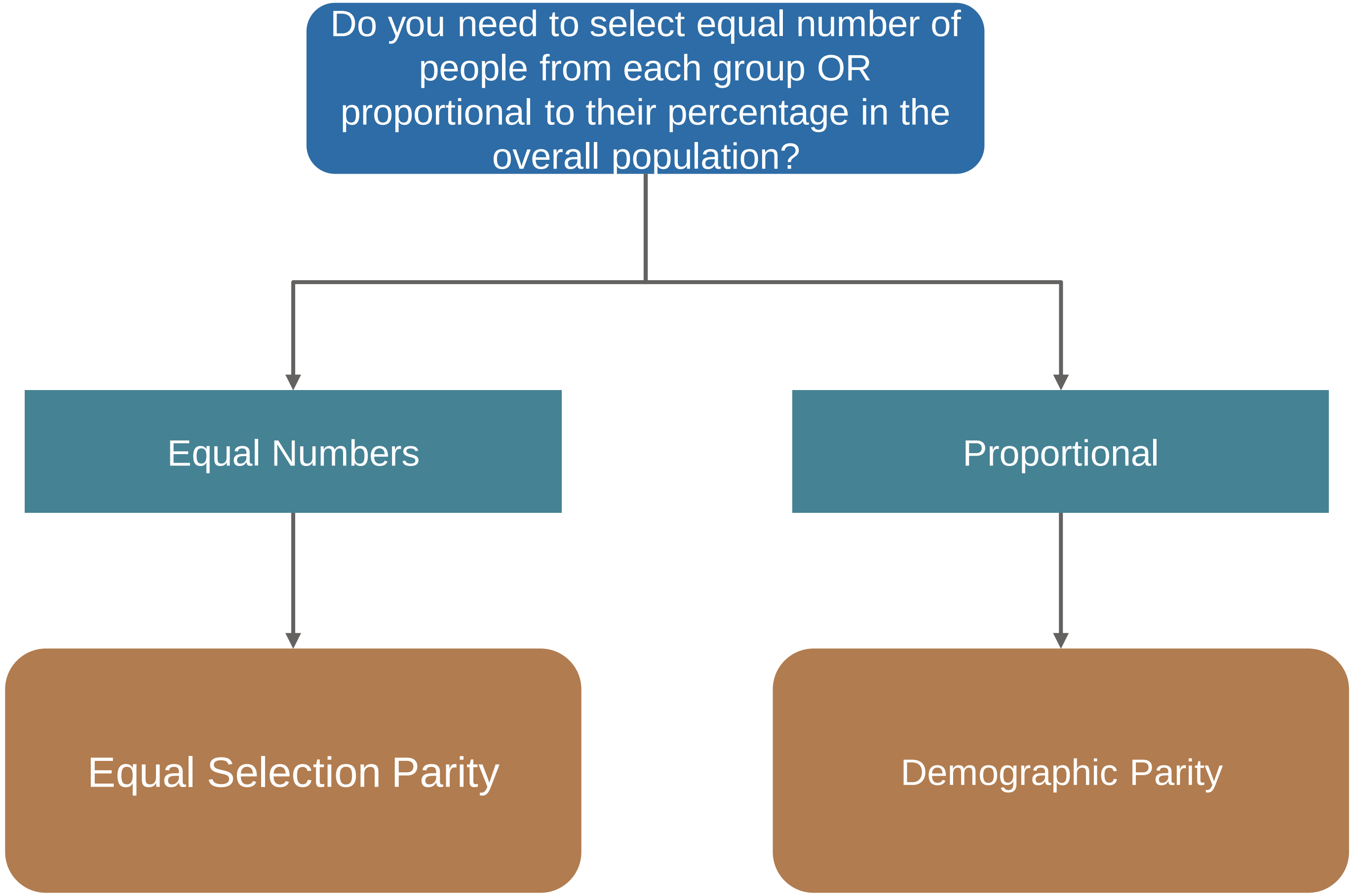
Which fairness metrics should we evaluate? http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

Choosing a fairness metric



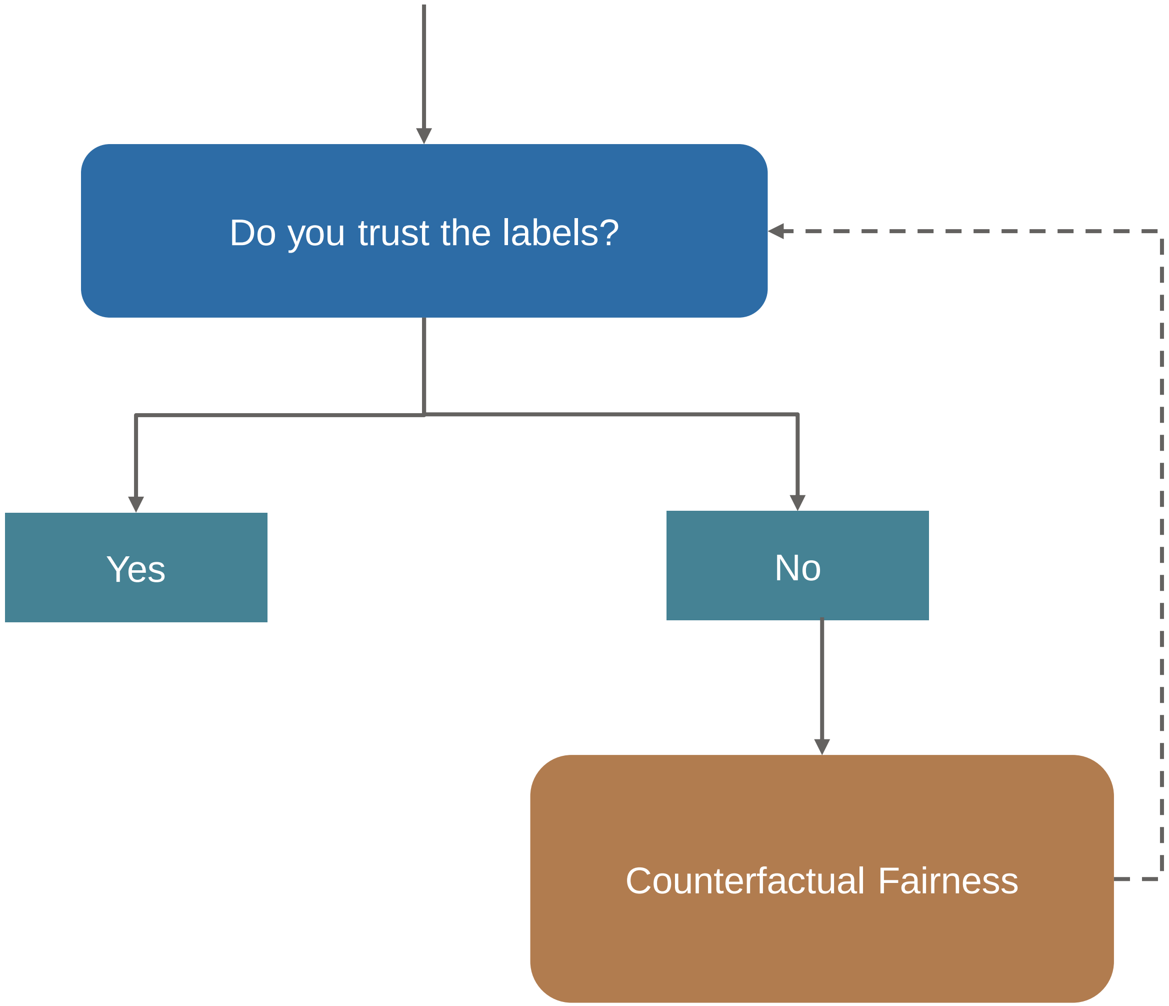
Source: 2019, Saleiro, Kuester, Hinkson, London, Stevens, Anisfeld, Rodolfa, Ghani. Aequitas_ A Bias and Fairness Audit Toolkit, arXiv Working Paper.

Choosing a fairness metric



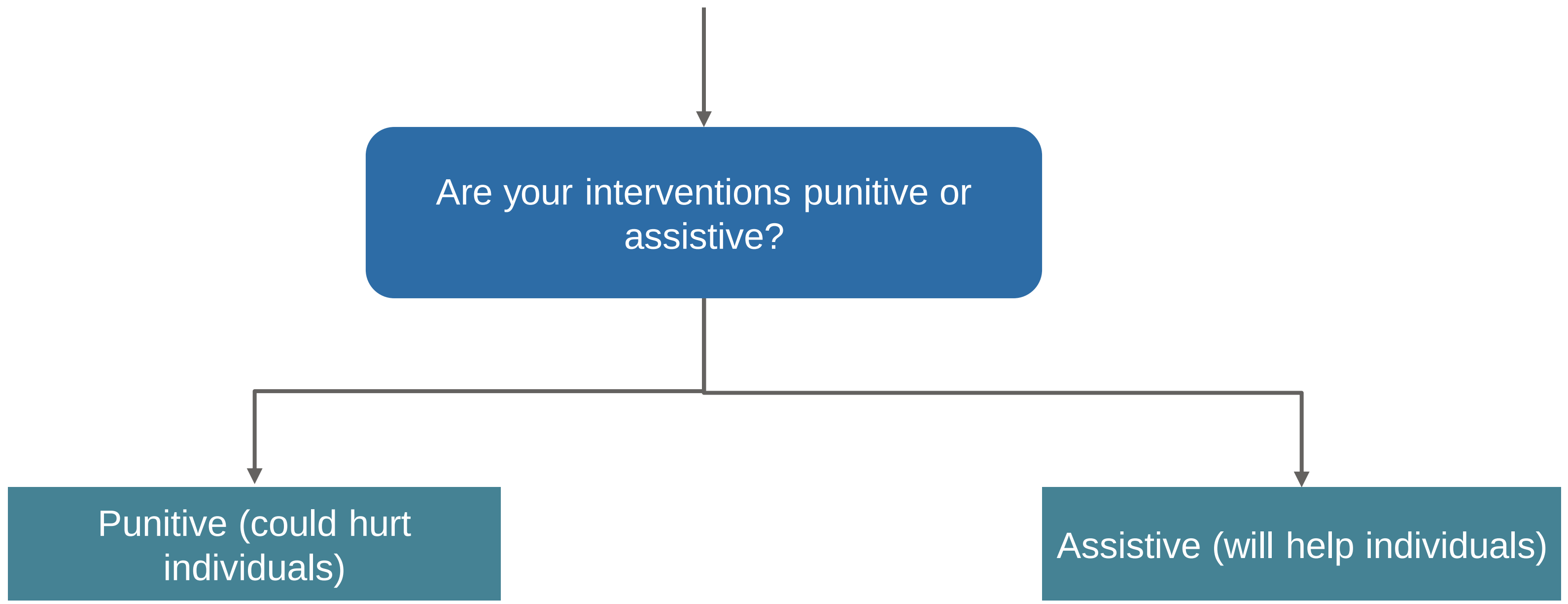
Source: 2019, Saleiro, Kuester, Hinkson, London, Stevens, Anisfeld, Rodolfa, Ghani. Aequitas_A Bias and Fairness Audit Toolkit, arXiv Working Paper.

Choosing a fairness metric



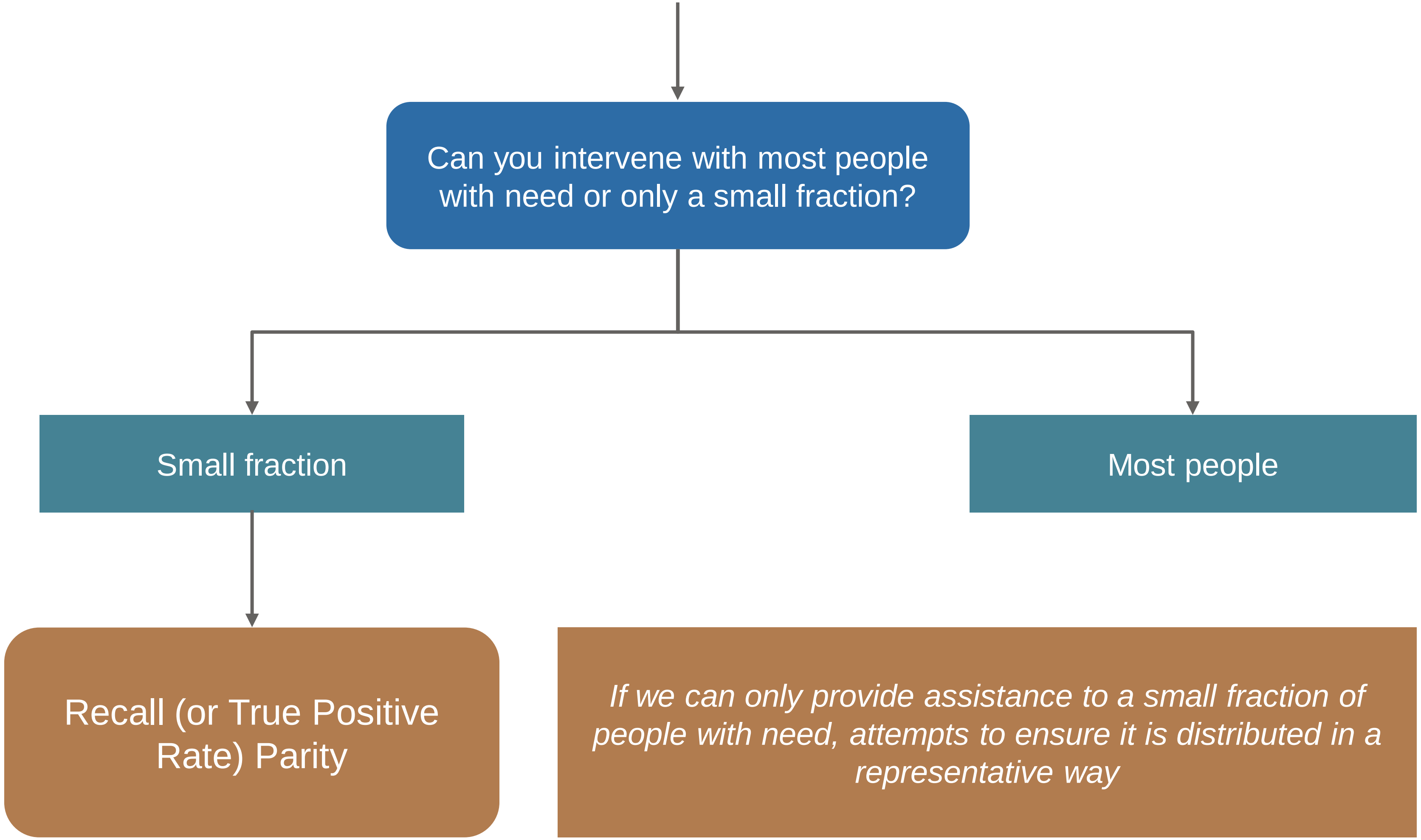
Source: 2019, Saleiro, Kuester, Hinkson, London, Stevens, Anisfeld, Rodolfa, Ghani. Aequitas_A Bias and Fairness Audit Toolkit, arXiv Working Paper.

Choosing a fairness metric



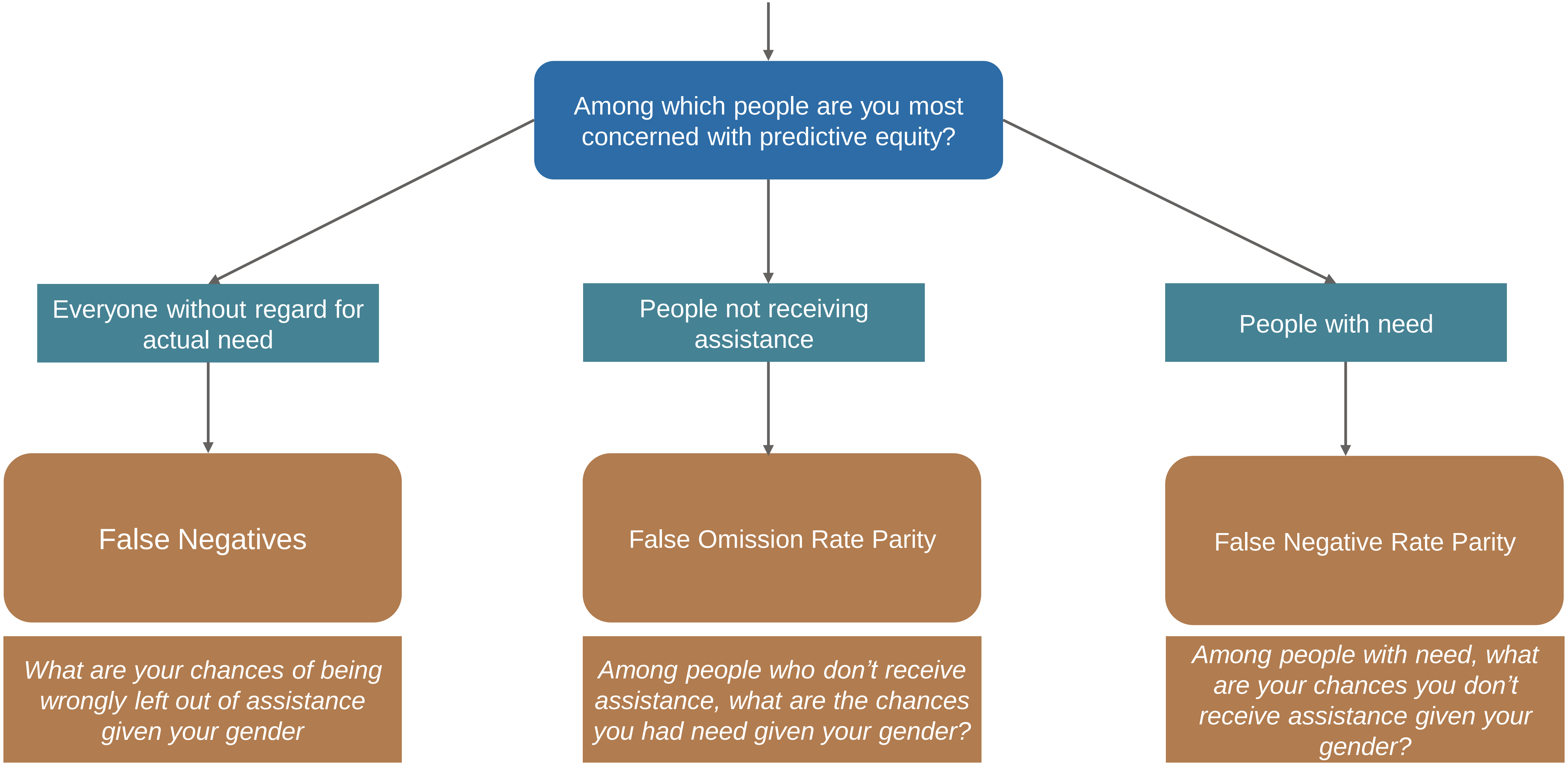
Source: 2019, Saleiro, Kuester, Hinkson, London, Stevens, Anisfeld, Rodolfa, Ghani. Aequitas_ A Bias and Fairness Audit Toolkit, arXiv Working Paper.

Choosing a fairness metric

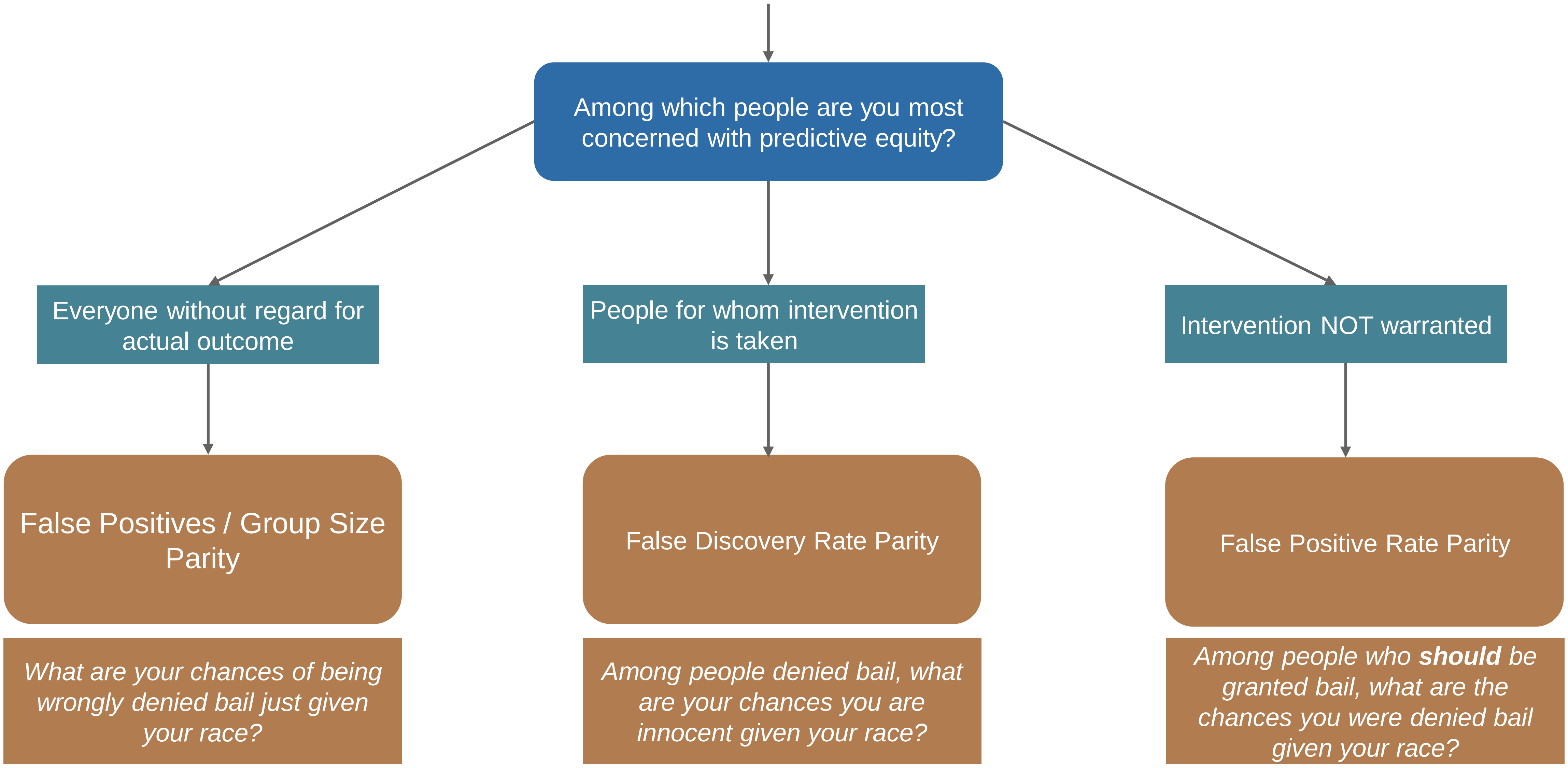


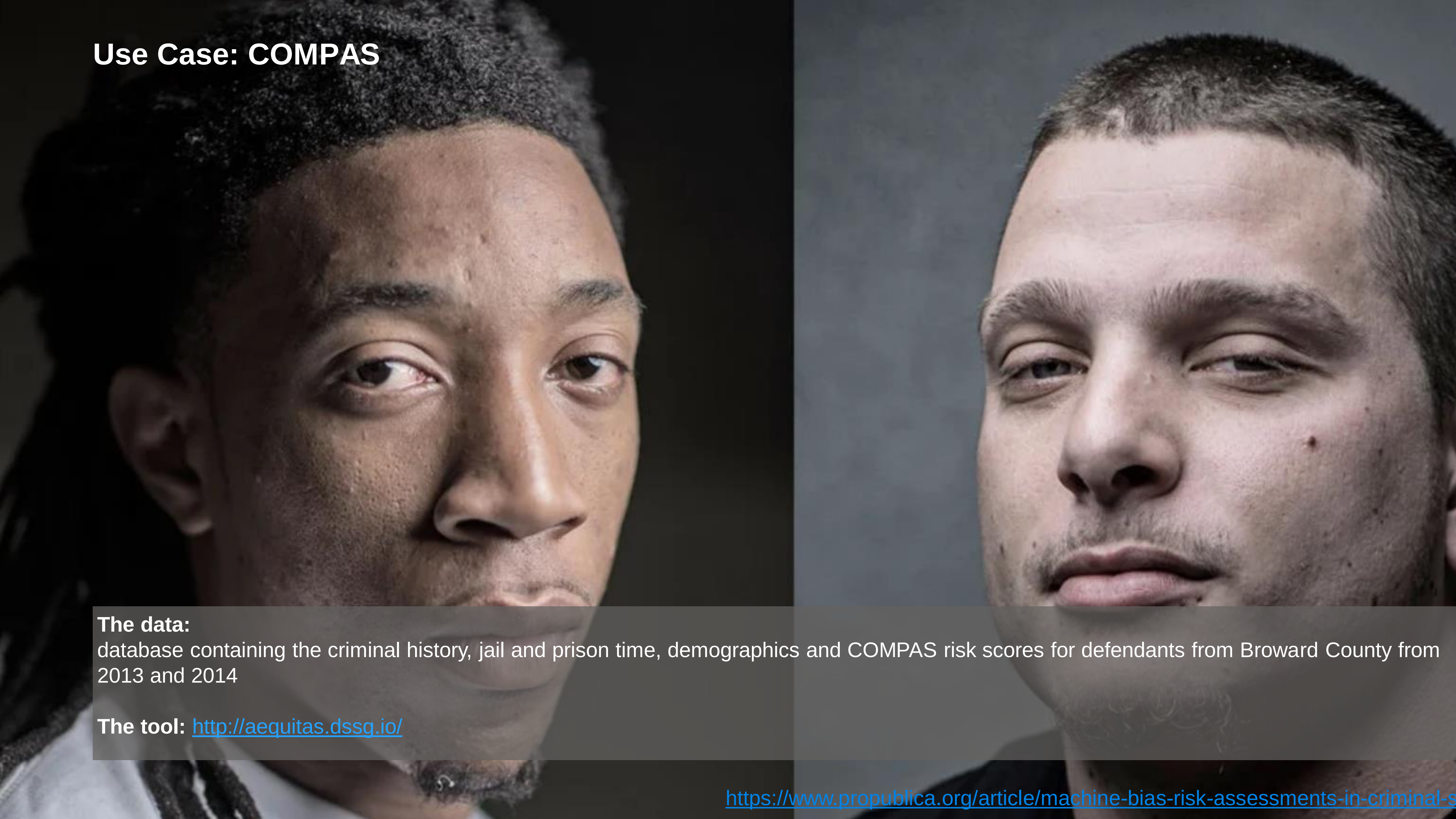
Source: 2019, Saleiro, Kuester, Hinkson, London, Stevens, Anisfeld, Rodolfa, Ghani. Aequitas_ A Bias and Fairness Audit Toolkit, arXiv Working Paper.

Choosing a fairness metric



Choosing a fairness metric





Use Case: COMPAS

The data:
database containing the criminal history, jail and prison time, demographics and COMPAS risk scores for defendants from Broward County from 2013 and 2014

The tool: http://aequitas.dssg.io/

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

Questions

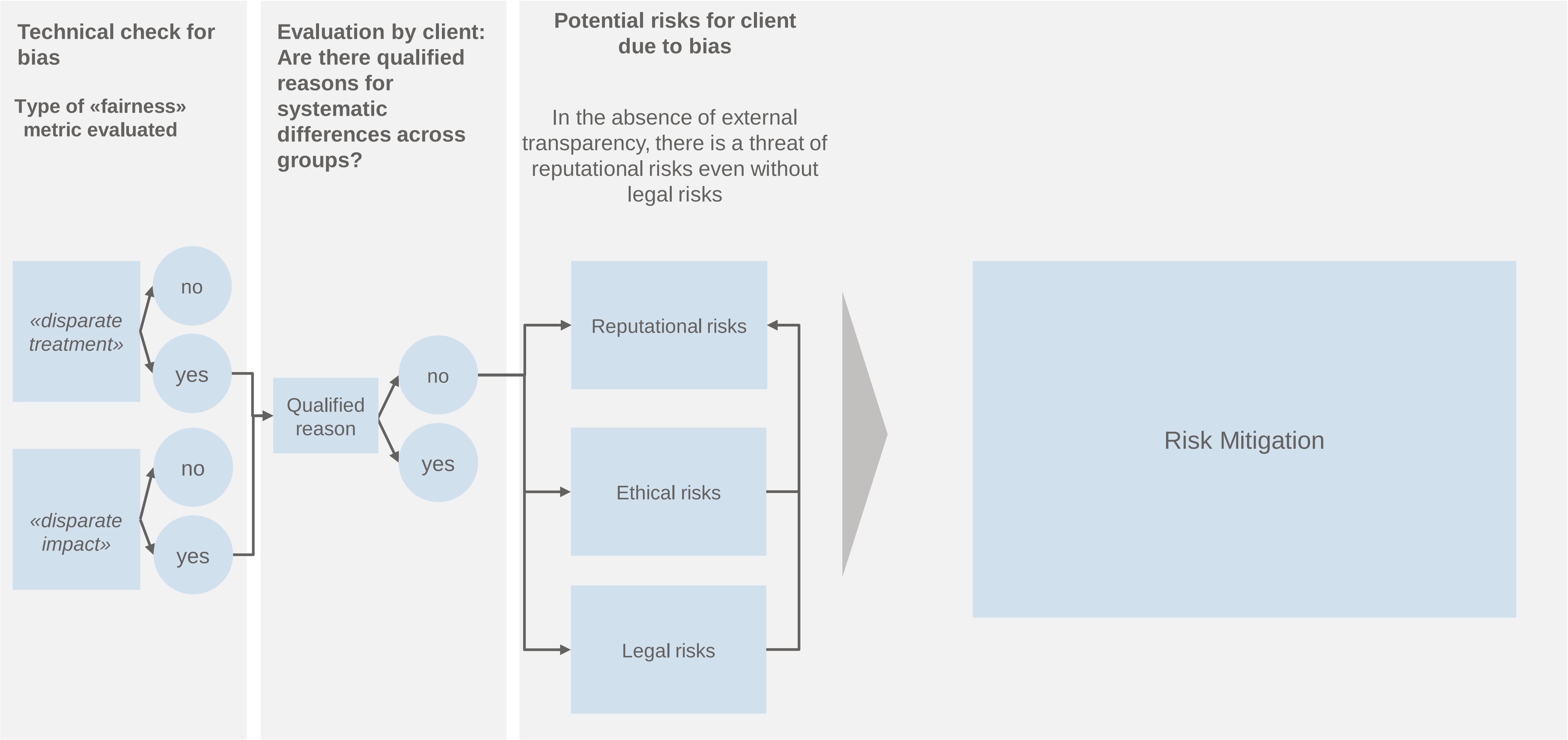
- **Based on which fairness criteria does ProRepublica get to their result?**

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

- **What happens if you set the disparity intolerance to 60%?**
- **Why do you think that the type of fairness criterion used is not mentioned in the article?**

https://colab.research.google.com/github/dssg/aequitas/blob/update_compas_notebook/docs/source/examples/compas_demo.ipynb#disparity_calc

What comes after the quantitative analysis?



Closing remarks: EU AI Act – Proposal for regulation

2. Training, validation and testing data sets shall be subject to appropriate data governance and management practices. Those practices shall concern in particular,

   (a) the relevant design choices;

   (b) data collection;

   (c) relevant data preparation processing operations, such as annotation, labelling, cleaning, enrichment and aggregation;

   (d) the formulation of relevant assumptions, notably with respect to the information that the data are supposed to measure and represent;

   (e) a prior assessment of the availability, quantity and suitability of the data sets that are needed;

   (f) examination in view of possible biases;

   (g) the identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed.

Sources

- ***Aequitas: A Bias and Fairness Audit Toolkit****,* Anisfeld, Ghani, Hinkson, Kuester, London, Rodolfa, Saleiro, Stevens, 2021, [online] Available at: https://arxiv.org/pdf/1811.05577.pdf [Accessed 9 August 2021].
- ***Automating Society Report 2020***, [ebook] Available at: https://automatingsociety.algorithmwatch.org/ [Accessed 9 August 2021].
- ***Bias Preservation in Machine Learning*****,** Wachter S., Mittelstadt, B., Russell, C. , 2021,West Virginia Law Review, https://ora.ox.ac.uk/objects/uuid:0c4cc51d-b2d3-4843-82ad-928e3b33e119 [Accessed 9 August 2021].
- ***Why Fairness Cannot Be Automated: Bridging the Gap Between the EU Non-Discrimination Law and AI*****,** Wachter S., Mittelstadt, B., Russell, C. 2021, Computer Law and Security Review. https://www.sciencedirect.com/science/article/abs/pii/S0267364921000406 [Accessed 9 August 2021].