4. Choosing a Fairness metric in practice

Which definition to choose?

- Not mathematically possible to construct an algorithm that simultaneously satisfies all reasonable definitions of a "fair" or "unbiased" algorithm [Kleinberg et al., 2016], [Chouldechova, 2017].

- Deciding which definition to use must be done in accordance with governance structures.

- Case-by-case basis. Often, context dictates whether the equality of opportunity or equality of outcome should be chosen.

Example situation: Facial Recognition

- No selection process that may require equal representativeness
- Ground-truth label is trusted

→ **Equality of Opportunity**

Example situation: Hiring

Algorithm selects top 100 candidates.

The problem: more white candidates selected in proportion (Disparate Impact = 0.7). But equality of opportunity metrics are good (Average odds difference).

- **Option 1:** Do nothing (Equality of opportunity) → darker skinned candidates complain that the data used to train the model contains bias.
- **Option 2:** Mitigate (Equality of Outcome) → a non-selected white candidate complains he was more qualified than a darker skinned candidates who was selected.

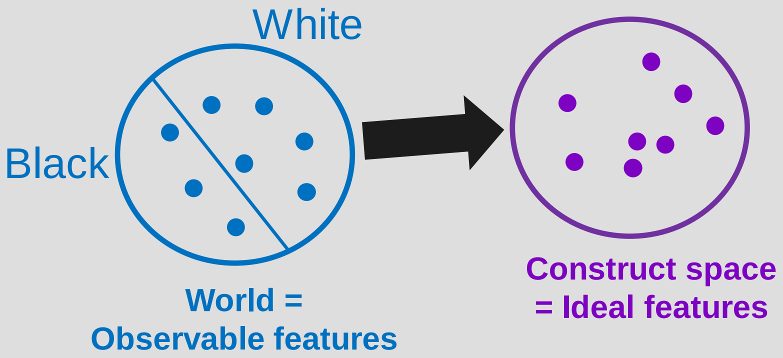
Which option to choose? How to justify a course of action?

Rule of thumb

- If you trust the ground-truth labels→ **Equality of Opportunity**
- If you don’t or if the label may contain bias too? **Equality of Outcome** or **Equality of Opportunity ?**

Two worldviews

[Friedler et al., 2016]
Idea of construct space



- Equality of Opportunity: What you see is what you get (WYSIWYG)

➔ construct space and observed space are essentially the same

- Equality of Outcome: We're all equal (WAE)

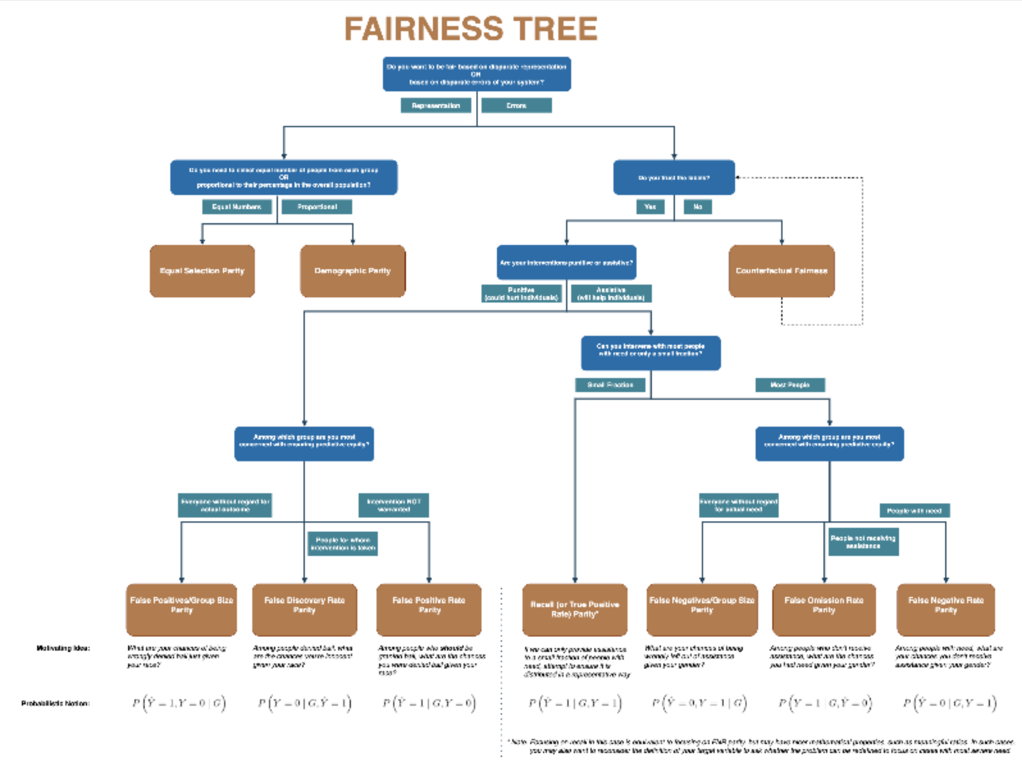
➔ Structural bias. Observed space not a good representation of construct space.

SAT scores

- **WYSIWYG:** worldview says that the score correlates well with future success and there is a way to use the score to correctly compare the abilities of applicants.

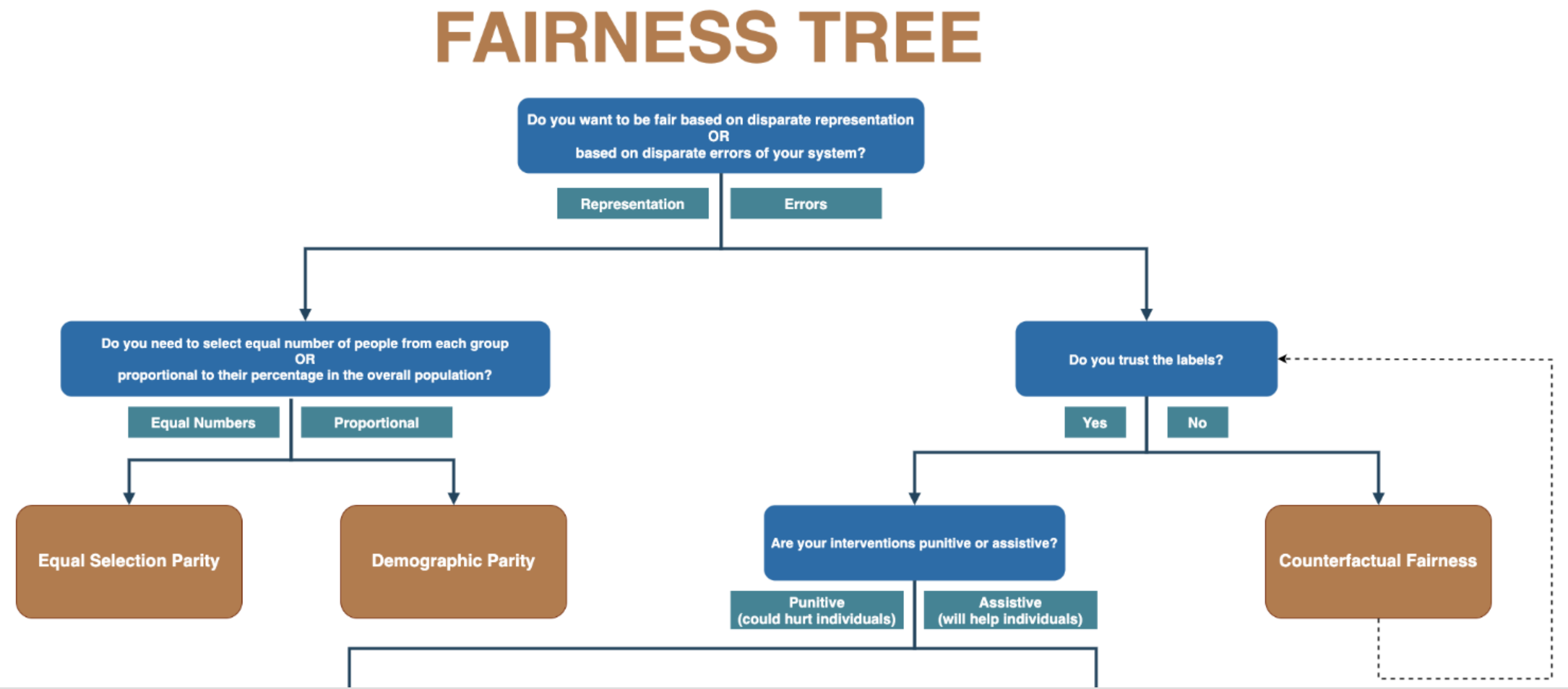
- **WAE:** worldview says that the SAT score may contain structural biases so its distribution being different across groups should not be mistaken for a difference in distribution in ability.

Fairness Tree



Aequitas

http://www.datasciencepublicpolicy.org/projects/aequitas/


FAIRNESS TREE
Do you want to be fair based on disparate representation
OR
based on disparate errors of your system?
Representation
Errors
Do you need to select equal number of people from each group
OR
proportional to their percentage in the overall population?
Equal Numbers
Proportional
Equal Selection Parity
Demographic Parity
Do you trust the labels?
Yes
No
Are your interventions punitive or assistive?
Punitive
(could hurt individuals)
Assistive
(will help individuals)
Counterfactual Fairness

Conclusion