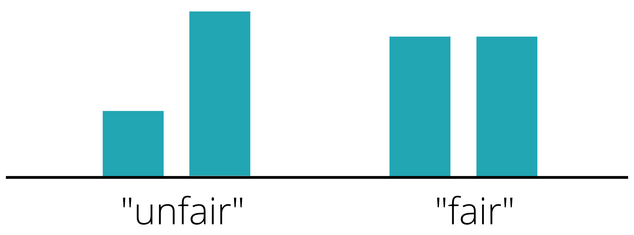
Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics

Joachim Baumann*, Corinna Hertweck*, Michele Loi and Christoph Heitz

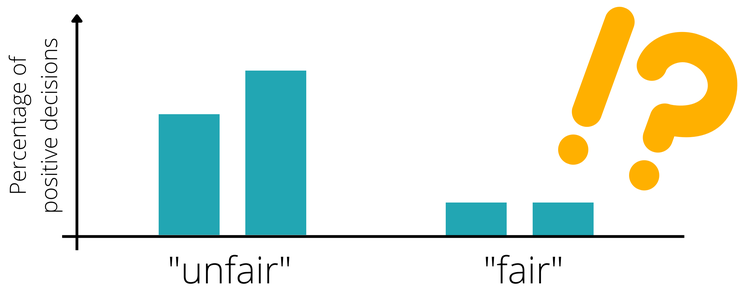
* Equal contribution

1. GROUP FAIRNESS: DEFINITION

Usually defined for binary decision making system
Compare socio-demographic groups: Equal wrt defined metric?
Examples: statistical parity, equality of opportunity, predictive parity



2. GROUP FAIRNESS: LIMITATIONS



- Leveling down: Equal values might make all groups worse off
- No consideration of consequences: Current criteria compare *decisions* instead of their *consequences*
- Difficult choice: Unclear how to choose between (incompatible) group fairness criteria, especially if none of them fit perfectly

Corinna Hertweck
https://hcorinna.github.io/
corinna.hertweck@zhaw.ch

Digital business card

3. MITIGATING THE LIMITATIONS: 3-STEP APPROACH

1. Utility function:
Define utility of decision depending on individual's other attributes
E.g., utility of loan approval positive if loan can be repaid, otherwise negative

2. Relevant groups:
What are the groups that have equal moral claims to utility, but are unlikely to receive equal utility?
E.g., all people who can repay their loan have equal claims to utility, but among those women and non-binary people are probably receiving less utility than men

3. Pattern of justice:
How should utility be distributed?
Here, theories of distributive justice come in, e.g.,

- **egalitarianism:** equalize utility between groups
- **maximin:** maximize utility of worst-off group
- **prioritarianism:** maximize overall utility, but give higher weight to worst-off group
- **sufficientarianism:** make sure that every group reaches a minimum level of utility

Extended definition of group fairness: **just** distribution of **utility** among **relevant groups**

4. RELATIONSHIP TO EXISTING GROUP FAIRNESS CRITERIA

Existing criteria can be mapped to our framework
⟶ Insights into assumptions of existing criteria about *utility function*, *relevant groups* and *pattern of justice*

Example: credit lending with *equality of opportunity* comparing women, men and non-binary people
Utility function: positive decision: utility of 1; negative decision: utility of 0
Relevant groups: Everyone who repays loan has same claim to utility, but women, men and non-binary people on average unlikely to receive same utility
Pattern of justice: egalitarianism

5. CONCLUSION

- Avoids limitations of current group fairness criteria
- Unique fairness criteria adapted to context
- Makes assumptions explicit

Zurich University of Applied Sciences zhaw School of Engineering
University of Zurich[UZH]
POLITECNICO MILANO 1863
FNSNF Swiss National Science Foundation