FAIRNESS COMPASS v1.0.0

Legend:
- ◇ Decision node
- ▭ Fairness definition
- ▭ Action node

START

Policy?
- Yes
- No

Equal base rates?
- Yes
- No
- No, but should be

Ground truth available?
- No
- Yes
- Eventually

Label annotation
- Failed
- Succeeded

Label bias?
- Yes
- No

Label correction
- Failed
- Succeeded

Type of representation?
- Equal numbers
- Proportional

Explaining variables?
- No
- Yes

Equal selection parity

Demographic parity

$P(\hat{Y} \mid A=0) = P(\hat{Y} \mid A=1)$

Conditional statistical parity

$P(\hat{Y}=1 \mid E=e, A=0) = P(\hat{Y}=1 \mid E=e, A=1)$

Evaluation metric most sensitive to fairness?
- Precision
- Recall

Output type? (Precision)
- Score
- Class label

Error types more sensitive to fairness?
- False negative + False positive
- False positive

Conditional use accuracy equality

$P(Y=y \mid \hat{Y}=y, A=0) = P(Y=y \mid \hat{Y}=y, A=1) \; \forall \; y \in \{0,1\}$

Predictive parity

$P(Y=1 \mid \hat{Y}=1, A=0) = P(Y=1 \mid \hat{Y}=1, A=1)$

Output type? (Recall)
- Class label
- Score

Which class is more sensitive to fairness?
- Positive class
- Negative class

Balance for positive class

$E[S \mid Y=1, A=0] = E[S \mid Y=1, A=1]$

Balance for negative class

$E[S \mid Y=0, A=0] = E[S \mid Y=0, A=1]$

Calibration

$P(Y=1 \mid S=s, A=0) = P(Y=1 \mid S=s, A=1) \; \forall \; s \in [0,1]$

Error types most sensitive to fairness?
- False negative + False positive
- False negative
- False positive

Equalized odds

$P(\hat{Y}=1 \mid Y=y, A=0) = P(\hat{Y}=1 \mid Y=y, A=1) \; \forall \; y \in \{0,1\}$

Equalized opportunities

$P(\hat{Y}=1 \mid Y=1, A=0) = P(\hat{Y}=1 \mid Y=1, A=1)$

Predictive equality

$P(\hat{Y}=1 \mid Y=0, A=0) = P(\hat{Y}=1 \mid Y=0, A=1)$