

BREAKING THE WALL OF INVISIBLE UNFAIRNESS IN MACHINE LEARNING RESEARCH

Yijun Bian

China

University of Copenhagen

the indicator function

model prediction on the raw instance

model prediction when only sensitive attribute(s) are changed

$$\ell_{\text{bias}}(f, \mathbf{x}) = \mathbb{I} \left(f(\tilde{\mathbf{x}}, \mathbf{a}) \neq f(\tilde{\mathbf{x}}, \tilde{\mathbf{a}}) \right)$$

non-sensitive attributes

sensitive attribute(s)

sensitive attribute(s) that are slightly disturbed

Algorithmic fairness: Not a purely technical
but socio-technical property

QUESTIONS?