Do Existing Fairness Measures Suffice? Assessing Discrimination in Algorithmic Decision-Making

Yijun Bian [#] University of Copenhagen, Denmark yibi@di.ku.dk
Yijun is on the job market! Scan me —>
Lei You Technical University of Denmark, Denmark
Yuya Sasaki The University of Osaka, Japan

TL;DR

Background Increasing reliance on AI/ML in high-stakes domains raises acute concerns about discrimination. Yet existing fairness metrics remain constrained with limited applicability, especially when facing complex scenarios beyond a protected attribute.

Brief We comprehensively analyse the insufficiency of existing fairness measures that are widely used, which hopefully provides interesting insights about them concerning non-binary cases.

References

Three statistical non-discrimination criteria Independence ($A \perp\!\!\!\perp R$), separation ($R \perp\!\!\!\perp A \mid Y$), and sufficiency ($Y \perp\!\!\!\perp A \mid R$).

Take *demographic parity (DP)* for example

$$\mathrm{DP}(f) = |\mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = 0) - \mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = 1)|, \tag{1a}$$

$$\mathrm{EOpp}(f) = |\mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = 0, y = 1) - \mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = 1, y = 1)|,$$

$$\mathrm{PP}(f) = |\mathbb{P}(y = 1 \mid a_1 = 0, f(\mathbf{x}, a_1) = 1) - \mathbb{P}(y = 1 \mid a_1 = 1, f(\mathbf{x}, a_1) = 1)|.$$

DP's possible extensions

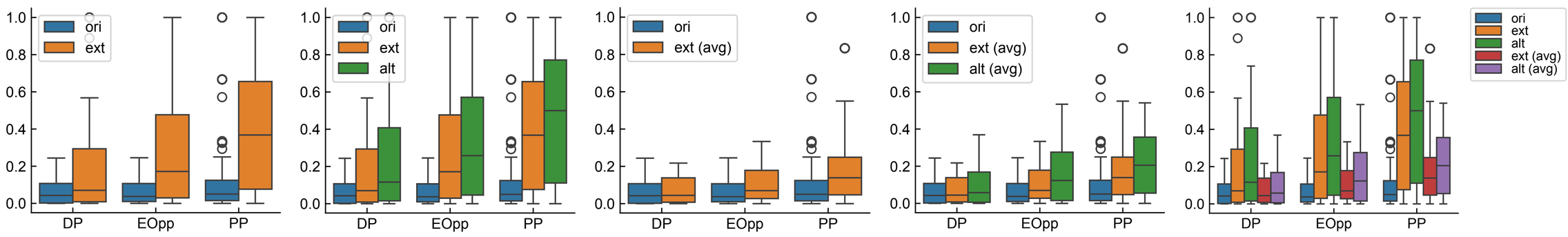
$$\mathrm{DP}'(f) = |\mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 \neq 1) - \mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = 1)|, \tag{2a}$$

$$\mathrm{DP}^{\mathrm{ext}}(f) = \max_{j \in \mathscr{A}_1} \{|\mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = j) - \mathbb{P}(f(\mathbf{x}, a_1) = 1)|\}, \tag{2b}$$

$$\mathrm{DP}^{\mathrm{alt}}(f) = \max_{j,k \in \mathscr{A}_1, k \neq j} \{|\mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = j) - \mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = k)|\}. \tag{2c}$$

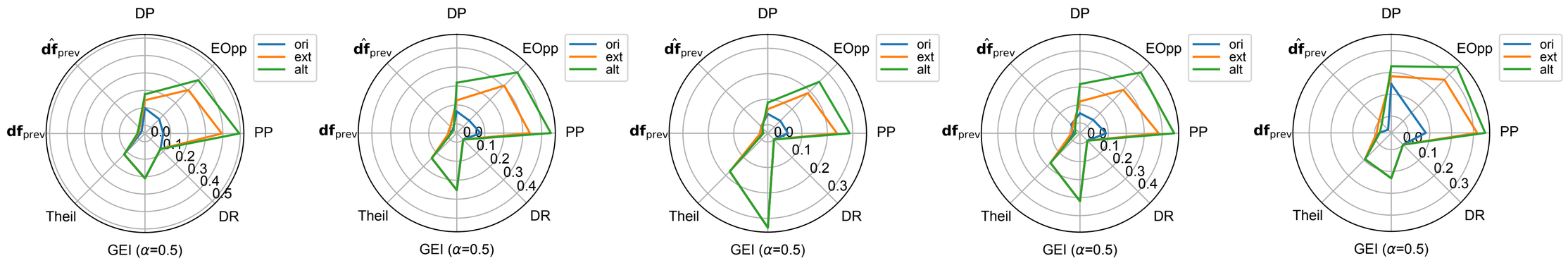
Empirical Analysis

Binarisation underestimates discrimination



(a) (b) (c) (d) (e)

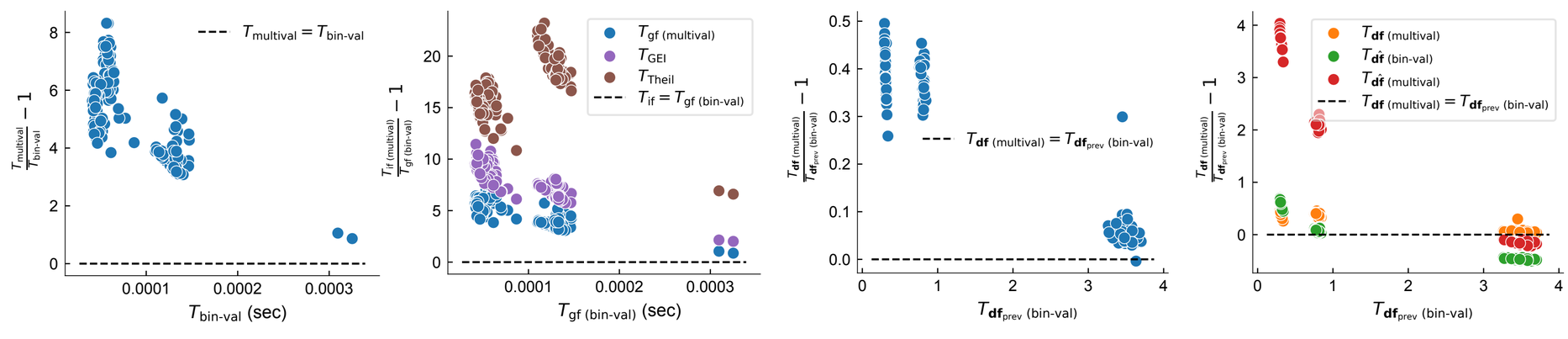
Figure 1: Comparison of three commonly used group fairness measures and their extensions.



(a) (b) (c) (d) (e)

Figure 2: This pattern recurs across datasets regardless of the learning algorithms in use, and suggests that it oversimplifies the structure of disadvantage and can systematically underestimate discrimination.

Traversal-based generalisation incurs computational cost



(a) (b) (c) (d)

Figure 3: Time cost comparison. (a–b) Group fairness, their extensions, and individual fairness; (c–d) HFM.

The 4th Deployable AI Workshop @ AAAI 2026, Singapore.