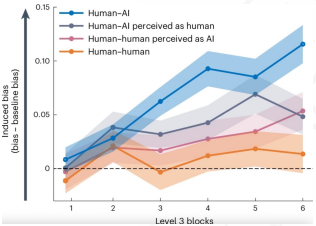
Do Existing FAIRNESS *Measures Suffice?*
Assessing Discrimination in Algorithmic Decision-Making

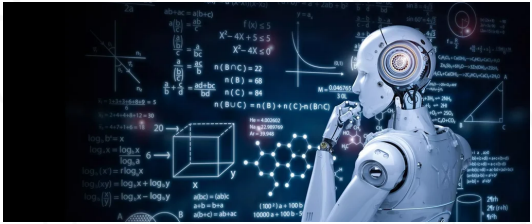
Yijun BIAN

Department of Computer Science
University of Copenhagen

27 August 2025

AI/ML is everywhere now





Social identity biases exist not only in human psychology and social behaviour, but also are present in artificial intelligence (AI) systems.[1]

When humans and AI interact, even minute perceptual, emotional and social biases[2]—originating either from AI systems or humans—leave human beliefs more biased, potentially forming a feedback loop.[3]

---

[1] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453. DOI: 10.1126/science.aax2342; Tiancheng Hu et al. "Generative language models exhibit social identity biases". In: *Nat Comput Sci* (2024), pp. 1–11; Richard J Chen et al. "Algorithmic fairness in artificial intelligence for medicine and healthcare". In: *Nat Biomed Eng* 7.6 (2023), pp. 719–742.

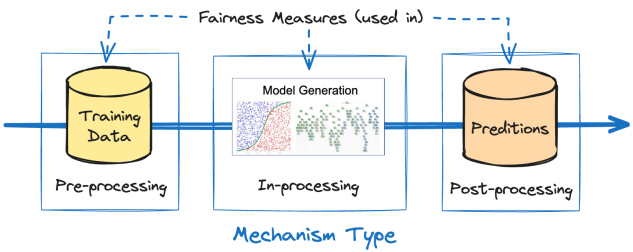
[2] Are Skeie Hermansen et al. "Immigrant–native pay gap driven by lack of access to high-paying jobs". In: *Nature* (2025), pp. 1–7.

[3] Moshe Glickman and Tali Sharot. "How human-AI feedback loops alter human perceptual, emotional and social judgements". In: *Nat Hum Behav* 9 (2025), pp. 345–359. DOI: 10.1038/s41562-024-02077-2; Madalina Vlasceanu and David M Amodio. "Propagation of societal gender inequality by internet search algorithms". In: *Proc Natl Acad Sci U.S.A.* 119.29 (2022), e2204529119.

Existing work about fairness

Many sources of bias [4]

Mechanisms to enhance fairness [5]



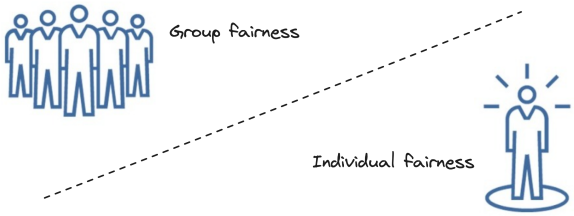
[4] *Unintential*: Limited and coarse features, sample size disparity (less data by definition about minority populations), skewed sample (feedback loops), tainted examples, features that act as proxies; *Intentional*: conscious prejudice.

[5] ***Pre-*** and ***post-processing*** *mechanisms* normally function by manipulating input or output, while ***inprocessing*** *mechanisms* introduce fairness constraints into training procedures or algorithmic objectives.

Existing work about fairness

Many sources of bias

Mechanisms to enhance fairness
Types of fairness measures [4,5]



Challenging: incompatibility,[6] multi-attribute protection, etc.

[4] ***Distributive*** *fairness*: group fairness, individual fairness, counterfactual fairness, etc.; ***Procedural*** *fairness*

[5] *Group fairness* focuses on statistical/demographic equality among groups defined by sensitive attributes, while *individual fairness* follows a principle that "similar individuals should be evaluated or treated similarly."

[6] Tensions between notions of fairness, between fairness and accuracy, between different methods for achieving fairness

Three statistical non-discrimination criteria

Statistical non-discrimination criteria are
properties of the joint distribution of a sensitive attribute (SA, aka. protected attribute) A, target variable y, the classifier $f(\cdot)$ or score R, sometimes including features X.

The three key criteria[7] are:

1. ***independence*** — Random variables (A, R) satisfy independence if $A \perp\!\!\!\perp R$
2. ***separation*** — Random variables (R, A, Y) satisfy separation if $R \perp\!\!\!\perp A \mid Y$
3. ***sufficiency*** — Random variables (R, A, Y) satisfy sufficiency if $Y \perp\!\!\!\perp A \mid R$

These criteria are rarely satisfied all at once, except in degenerate cases.[8]

[7] In essence, the latter two require the same recall or precision for each group, respectively.

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. Cambridge, MA, USA: MIT Press, 2023; Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning*. fairmlbook.org, 2019; Reuben Binns. "Fairness in machine learning: Lessons from political philosophy". In: *FAT*. PMLR. 2018, pp. 149–159.

#1 Independence

Random variables (A, R) satisfy independence if $A \perp\!\!\!\perp R$

#1 statistical non-discrimination criterion

Demographic parity (DP),[9] aka. statistical parity[10]

$$\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = 0) = \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = 1). \tag{1}$$

[9] Pratik Gajane and Mykola Pechenizkiy. "On formalizing fairness in prediction with machine learning". In: *FAT/ML*. 2018; Ray Jiang et al. "Wasserstein fair classification". In: *UAI*. PMLR. 2020, pp. 862–872.

[10] Cynthia Dwork et al. "Fairness through awareness". In: *ITCS*. ITCS '12. Cambridge, Massachusetts: ACM, 2012, pp. 214–226. ISBN: 9781450311151; Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big Data* 5.2 (2017), pp. 153–163.

#1 Independence

Disparate impact (i.e., "80% rule"):[11] the larger the better

$$\frac{\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = 0)}{\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = 1)} \leqslant \tau = 0.8\,. \tag{2}$$

Disparate treatment[12]

$$\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1) = \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1)\,,\ a_1 \in \{0, 1\}\,. \tag{3}$$

It is also indicated as "**statistical parity** (SP)" in the literature,[13] that is,

$$\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid \ a_1 = j\ ) = \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1)\,,\ \forall j \in \mathcal{A}_1 = \{1, 2, ..., n_{a_1}\}\,.$$

[11] Michael Feldman et al. "Certifying and removing disparate impact". In: *SIGKDD*. 2015, pp. 259–268; Muhammad Bilal Zafar et al. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment". In: *WWW*. 2017, pp. 1171–1180.

[12] Zafar et al., see n. 11; Christian Haas. "The price of fairness - A framework to explore trade-offs in algorithmic fairness". In: *ICIS*. Association for Information Systems. 2019.

[13] Haas, see n. 12; Sam Corbett-Davies et al. "Algorithmic decision making and the cost of fairness". In: *SIGKDD*. New York, NY, USA: ACM, 2017, pp. 797–806. ISBN: 9781450348874; Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. "Fair regression: Quantitative definitions and reduction-based algorithms". In: *ICML*. vol. 97. PMLR. 2019, pp. 120–129.

#2 Separation

Random variables (R, A, Y) satisfy separation if $R \perp\!\!\!\perp A \mid Y$

#2 statistical non-discrimination criterion

Equalised odds (EO)[14]

$$\mathbb{P}(f(\check{\mathbf{x}}, a_1) = 1 \mid a_1 = 0, y) = \mathbb{P}(f(\check{\mathbf{x}}, a_1) = 1 \mid a_1 = 1, y)\,,\ y \in \{0, 1\}\,. \tag{4}$$

Equality of opportunity (EOpp, or EO),[15] aka. treatment equality[16]

$$\mathbb{P}(f(\check{\mathbf{x}}, a_1) = 1 \mid a_1 = 0\,,\ y = 1) = \mathbb{P}(f(\check{\mathbf{x}}, a_1) = 1 \mid a_1 = 1\,,\ y = 1)\,. \tag{5}$$

[14] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of opportunity in supervised learning". In: *NIPS*. vol. 29. Red Hook, NY, USA, 2016, pp. 3323–3331; Haas, see n. 12.

[15] Hardt, Price, and Srebro, see n. 14; Gajane and Pechenizkiy, see n. 9; Haas, see n. 12.

[16] Richard Berk et al. "Fairness in criminal justice risk assessments: The state of the art". In: *Sociol Methods Res* 50.1 (2021), pp. 3–44.

#3 Sufficiency

Random variables (R, A, Y) satisfy sufficiency if $Y \perp\!\!\!\perp A \mid R$

#3 statistical non-discrimination criterion

Predictive parity (PP)[17]

$$\mathbb{P}(y = 1 \mid a_1 = 0, f(\check{\boldsymbol{x}}, a_1) = 1) = \mathbb{P}(y = 1 \mid a_1 = 1, f(\check{\boldsymbol{x}}, a_1) = 1). \tag{6}$$

Calibration[18] (concept). For two binary predictors $h_1, h_0 : \mathbb{R}^{n_d+1} \mapsto [0,1]$, h_1 classifies samples with $a_1 = 1$ and h_0 does samples with $a_1 = 0$. Any h_t ($t \in \{1,0\}$) is perfectly calibrated if

$$\forall p \in [0,1],\ \mathbb{P}_{(\check{\boldsymbol{x}}, a_i = t, y)}(y = 1 \mid h_t(\check{\boldsymbol{x}}, a_1) = p) = p.$$

It intuitively prevents the probability scores from carrying group-specific information.

[17] Chouldechova, see n. 10; Sahil Verma and Julia Rubin. "Fairness definitions explained". In: *FairWare*. 2018, pp. 1–7.

[18] Geoff Pleiss et al. "On fairness and calibration". In: *NIPS*. vol. 30. 2017; Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores". In: *ITCS*. 2017.

Individual fairness

Lipschitz condition[19,20]

A mapping/predictor $h\colon \mathcal{X} \times \mathcal{A}_1 = \mathcal{X} \times \{0,1\} \mapsto [0,1]$ satisfies the λ-Lipschitz property if for any $(\check{\boldsymbol{x}}, a_1), (\check{\boldsymbol{x}}', a_1')$,

$$\boldsymbol{d}_y(\, h(\check{\boldsymbol{x}}, a_1) \,, h(\check{\boldsymbol{x}}', a_1')) \leqslant \lambda \cdot \boldsymbol{d}_x(\, (\check{\boldsymbol{x}}, a_1) \,, (\check{\boldsymbol{x}}', a_1'))\,, \tag{7}$$

where $\boldsymbol{d}_y$ and $\boldsymbol{d}_x$ are (task-specific) distance metrics. Note that λ is a positive constant.

It can also be written as the probability Lipschitzness, i.e., $\mathbb{P}\left(\frac{\boldsymbol{d}_y(h(\check{\boldsymbol{x}}, a_1), h(\check{\boldsymbol{x}}', a_1'))}{\boldsymbol{d}_x((\check{\boldsymbol{x}}, a_1), (\check{\boldsymbol{x}}', a_1'))} \geqslant \epsilon\right) \leqslant \delta$; or the $(\epsilon - \delta)$ language formulation: $\boldsymbol{d}_x((\check{\boldsymbol{x}}, a_1), (\check{\boldsymbol{x}}', a_1')) \leqslant \epsilon \Rightarrow \boldsymbol{d}_y(h(\check{\boldsymbol{x}}, a_1), h(\check{\boldsymbol{x}}', a_1')) \leqslant \delta$, where $\epsilon \geqslant 0$ and $\delta \geqslant 0$.

In essence, individual fairness follows the principle that "similar individuals should be evaluated or treated similarly." A careful choice of distance metrics is crucial in ensuring fairness.[21]

[19] Dwork et al., see n. 10.

[20] Additionally, a predictor satisfies individual fairness (Gajane and Pechenizkiy, see n. 9) iff: $h(\check{\boldsymbol{x}}, a_1) \approx h(\check{\boldsymbol{x}}', a_1') \mid \boldsymbol{d}_x((\check{\boldsymbol{x}}, a_1), (\check{\boldsymbol{x}}', a_1')) \approx 0$, where $\mathcal{X}_a \triangleq \mathcal{X} \times \mathcal{A}$ and $\boldsymbol{d}_x : \mathcal{X}_a \times \mathcal{X}_a \mapsto \mathbb{R}$ is a distance metric for individuals.

[21] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. "k-NN as an implementation of situation testing for discrimination discovery and prevention". In: *SIGKDD*. 2011, pp. 502–510; Laura Boeschoten et al. "Achieving fair inference using error-prone outcomes". In: *Int J Interact Multimed Artif Intell* 6.5 (2021).

Individual fairness

General entropy indices[22] and the **Theil index**[23]

For a constant $\alpha \notin \{0, 1\}$, the generalised entropy indices for a problem with n instances are defined, to quantify algorithmic unfairness, as

$$\mathrm{GEI}^{\alpha} = \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} \left(\left(\frac{b_i}{\mu} \right)^{\alpha} - 1 \right), \tag{8}$$

where benefits $b_i = f(\check{\boldsymbol{x}}_i, a_{1i}) - y_i + 1$ and $\mu = \Sigma_i \, b_i / n$.

The Theil index is a special case for $\alpha = 1$, that is,

$$\mathrm{Theil} = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i}{\mu} \log \left(\frac{b_i}{\mu} \right). \tag{9}$$

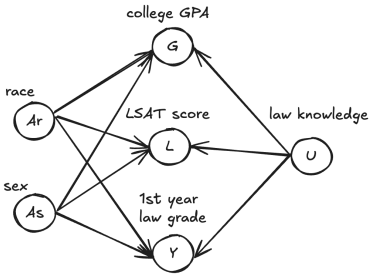
They are used additionally to group fairness measures to compare different algorithms and determine which one is considered the fairest from an individual perspective.

[22] Till Speicher et al. "A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices". In: *SIGKDD*. 2018, pp. 2239–2248.

[23] Haas, see n. 12.

Individual fairness

Example/Illustration:
Law school success



As	Ar	G	L	Y	U
male	black	👎	👍	👎	(👍)

1. Computing unobserved variables in causal model
2. Change A (that is, sensitive attribute(s))
3. Recompute observed variables in causal model

classifier
hat{Y}(G,L)

As	Ar ← a'	G	L	Y	U
male	white	👎	👍	👎	👍

As	Ar ←a'	$G_{[Ar \leftarrow a']}$	$L_{[Ar \leftarrow a']}$	$Y_{[Ar \leftarrow a']}$	$U_{[Ar \leftarrow a']}$
male	white	👍	👍	👍	👍

Counterfactual fairness[24] (definition). A predictor $\hat{Y}$ is *counterfactually fair* if given observations $\mathcal{X} = \check{\boldsymbol{x}}$ and $A = \boldsymbol{a}$ we have that, $\mathbb{P}(\hat{Y}_{A \leftarrow \boldsymbol{a}} = y \mid \mathcal{X} = \check{\boldsymbol{x}}, A = \boldsymbol{a}) = \mathbb{P}(\hat{Y}_{A \leftarrow \boldsymbol{a}'} = y \mid \mathcal{X} = \check{\boldsymbol{x}}, A = \boldsymbol{a})$, for all y and $\boldsymbol{a}' \neq \boldsymbol{a}$.
Proxy discrimination [25]. A predictor $\hat{Y}$ exhibits no *individual proxy discrimination* based on a proxy P if for all $\boldsymbol{p}, \boldsymbol{p}'$, we have $\mathbb{P}(\hat{Y} \mid \mathrm{do}(P = \boldsymbol{p}))= \mathbb{P}(\hat{Y} \mid \mathrm{do}(P = \boldsymbol{p}'))$, where $\mathrm{do}(P = \boldsymbol{p})$ denotes an intervention on P.

[24] Matt J Kusner et al. "Counterfactual fairness". In: *NIPS*. vol. 30. NIPS Proceedings. 2017, pp. 4069–4079.

[25] (Niki Kilbertus et al. "Avoiding discrimination through causal reasoning". In: *NIPS*. vol. 30. 2017) Visually, intervening on P amounts to removing all incoming arrows of P in the graph; algebraically, it consists of replacing the structural equation of P by $P = \mathbf{p}$, i.e. we put point mass on the value $\boldsymbol{p}$.

Brief summary

Table 1: Summary of existing fairness measures.

Name of *measure*	Fairness type[1]	Meaning		Applicable situation(s) in definition			Non-binary handling[4]	
		quant.[2]	fairer	#label (n_c)	#sen-att (n_a)[3]	#values per $\mathcal{A}_i$	$n_{a_i} > 2$	$n_a > 1$
Demographic parity (aka. statistical parity)	*, group-	yes	lower value	binary	singular	bi-valued	yes	indirectly
Disparate impact /80% rule	*, group-	yes	*larger* value	binary	singular	bi-valued	yes	indirectly
Disparate treatment	*, group-	poss.	lower value	binary	singular	bi- (multi- allowed)	yes	indirectly
Conditional statistical parity	*, group-	poss.	lower value	binary	singular	multi-valued	—	indirectly
Bounded group loss	*, group-	poss.	lower value	binary	singular	multi-valued	—	indirectly
Strategic minimax fairness	*, group-	no	—	bi-/multi-	singular	multi-valued	—	indirectly
Equalised odds	*, group-	yes	lower value	binary	singular	bi-valued	yes	indirectly
Equality of opportunity	*, group-	yes	lower value	binary	singular	bi-valued	yes	indirectly
Predictive equality	*, group-	poss.	lower value	binary	singular	multi-valued	—	indirectly
γ-subgroup fairness	*, group-	yes	lower value	binary	singular	bi-valued	yes	indirectly
Predictive parity	*, group-	yes	lower value	binary	singular	bi-valued	yes	indirectly
Lipschitz condition	*, individual-	no	—	binary	singular	bi-valued	yes	indirectly
General entropy indices (and the Theil index)	*, individual-	yes	lower value	binary	singular	multi-valued	—	indirectly
Counterfactual fairness	*, individual-	no	—	binary	allows plural	multi- allowed	yes	indirectly
Proxy discrimination	*, individual-	no	—	binary	singular	multi- allowed	yes	indirectly
Discriminative risk[26]	*,[5]	yes	lower value	bi-/multi-	allows plural	multi-valued	—	—
Harmonic fairness via manifold[27]	*,[5]	yes	lower value	bi-/multi-	allows plural	multi-valued	—	—
Multiaccuracy	*,group-	poss.	lower value	binary	singular	multi-valued	—	indirectly
Differentially fair	*,group-	poss.	—	binary	allows plural	bi- (multi- allowed)	yes	indirectly
Group benefit ratio and worst-case min-max ratio	*,group-	yes	*larger* value	binary	allows plural	bi- (multi- allowed)	yes	indirectly
Feature-apriori fairness	procedural	yes	—	binary	—	—	yes	yes
Feature-accuracy fairness	procedural	yes	—	binary	—	—	yes	yes
Feature-disparity fairness	procedural	yes	—	binary	—	—	yes	yes
FAE-based procedural fairness	procedural	yes	lower value	binary	singular	bi-valued	yes	indirectly

[26] Yijun Bian and Kun Zhang. "Increasing fairness via combination with learning guarantees". In: *arXiv preprint arXiv:2301.10813* (2023). Under review.

[27] Yijun Bian and Yujie Luo. "Does machine bring in extra bias in learning? Approximating fairness in models promptly". In: *arXiv preprint arXiv:2405.09251* (2024). Under review; Yijun Bian, Yujie Luo, and Ping Xu. "Approximating discrimination within models when faced with several non-binary sensitive attributes". In: *arXiv preprint arXiv:2408.06099* (2024). Under review.

Binarisation underestimates discrimination

Statistical parity (SP)[28]

A predictor h satisfies *statistical parity* under a distribution over (X, A, Y) if $h(\boldsymbol{x})$ is independent of the protected attribute $\boldsymbol{a}$. Since $h(\boldsymbol{x}) \in [0, 1]$, this is equivalent to

$$\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) \geqslant z \mid a_1 = j) = \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) \geqslant z) \tag{10}$$

for all $j \in \mathcal{A}_1 = \{1, 2, ..., n_{a_1}\}$ and $z \in [0, 1]$.

Demographic parity (DP) $\quad |\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 \neq 1) - \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = 1)| \leqslant \varepsilon$

#1 statistical non-discrimination criterion

DP's extension and alternative[29] form

$$\max_{j \in \mathcal{A}_1} |\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = j) - \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1)|, \tag{11a}$$

$$\max_{j,k \in \mathcal{A}_1, j \neq k} |\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = j) - \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = k)|. \tag{11b}$$

[28] Corbett-Davies et al., see n. 13; Agarwal, Dudík, and Wu, see n. 13.
[29] Jiang et al., see n. 9.

Binarisation underestimates discrimination

Demographic parity (DP) $|\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 \neq 1) - \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = 1)| \leqslant \varepsilon$

Equality of opportunity (EOpp) $|\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 \neq 1, y = 1) - \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = 1, y = 1)| \leqslant \varepsilon$

Predictive parity (PP) $|\mathbb{P}(y = 1 \mid a_1 \neq 1, f(\check{\boldsymbol{x}}, a_1) = 1) - \mathbb{P}(y = 1 \mid a_1 = 1, f(\check{\boldsymbol{x}}, a_1) = 1)| \leqslant \varepsilon$

Three statistical non-discrimination criteria

EOpp's extension and alternative form

$$\max_{j \in \mathcal{A}_1} |\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = j, y = 1) - \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid y = 1)|, \tag{10a}$$

$$\max_{j,k \in \mathcal{A}_1, j \neq k} |\mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = j, y = 1) - \mathbb{P}(f(\check{\boldsymbol{x}}, a_1) = 1 \mid a_1 = k, y = 1)|. \tag{10b}$$

PP's extension and alternative form

$$\max_{j \in \mathcal{A}_1} |\mathbb{P}(y = 1 \mid a_1 = j, f(\check{\boldsymbol{x}}, a_1) = 1) - \mathbb{P}(y = 1 \mid f(\check{\boldsymbol{x}}, a_1) = 1)|, \tag{11a}$$

$$\max_{j,k \in \mathcal{A}_1, j \neq k} |\mathbb{P}(y = 1 \mid a_1 = j, f(\check{\boldsymbol{x}}, a_1) = 1) - \mathbb{P}(y = 1 \mid a_1 = k, f(\check{\boldsymbol{x}}, a_1) = 1)|. \tag{11b}$$

Binarisation underestimates discrimination

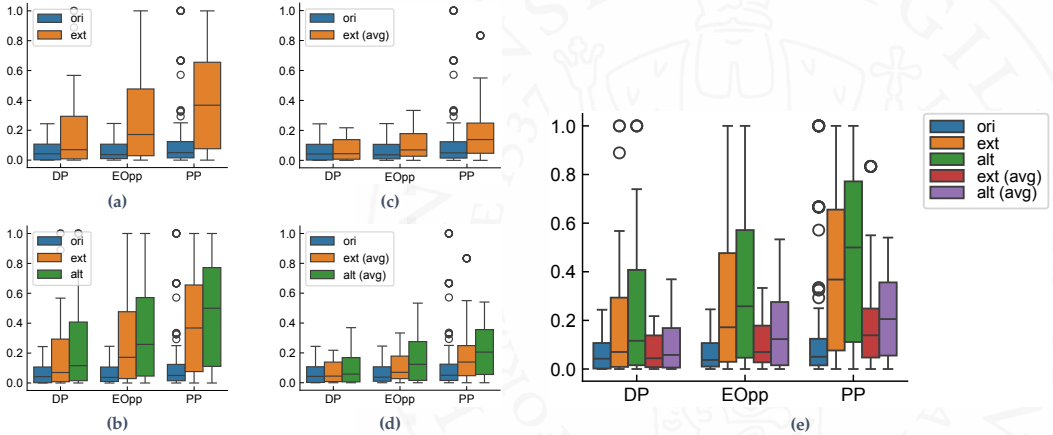


Figure 1: Comparison of three commonly used group fairness measures and their extensions, on Income, Compas PPR, and Compas PPVR datasets.
(a–b) Comparison between binarisation and the two extension forms, analogously to Eq. (11a) and (11b); note that binarisation is equivalent to their original definitions like (1). (c–d) Comparison between binarisation and their corresponding average forms. (e) Comparison between binarisation and all four extension formulas.

Binarisation underestimates discrimination

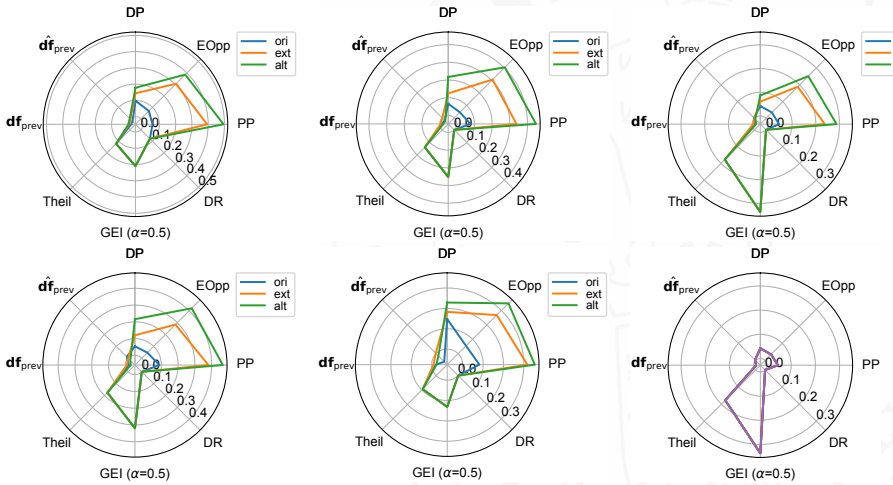


Figure 1: Comparison of fairness measures between their original definisions and their corresponding extension forms, on the Income dataset. (a–e) Using bagging, AdaBoost, LightGBM, AdaFair (trained using #1 sen-att), and AdaFair (trained using #2 sen-att), respectively.

Binarisation underestimates discrimination

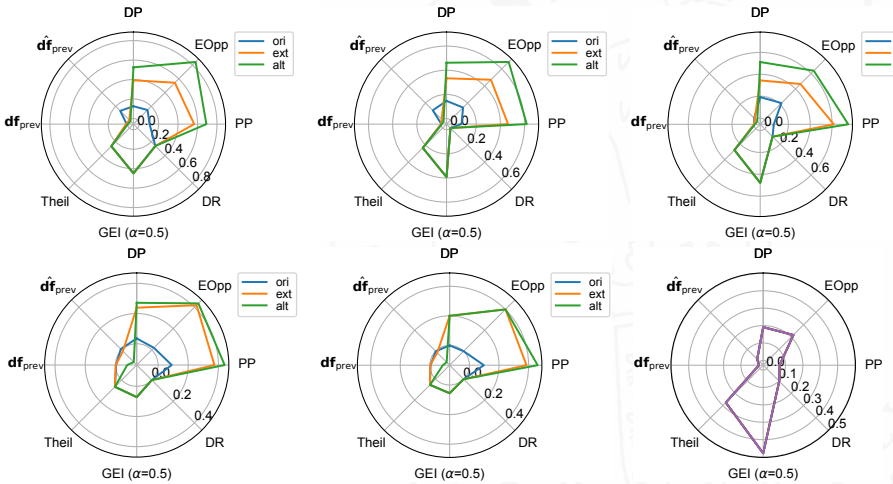


Figure 1: Comparison of fairness measures between their original definisions and their corresponding extension forms, on the Compas PPR dataset. (a–e) Using bagging, AdaBoost, LightGBM, AdaFair (trained using #1 sen-att), and AdaFair (trained using #2 sen-att), respectively.

Binarisation underestimates discrimination

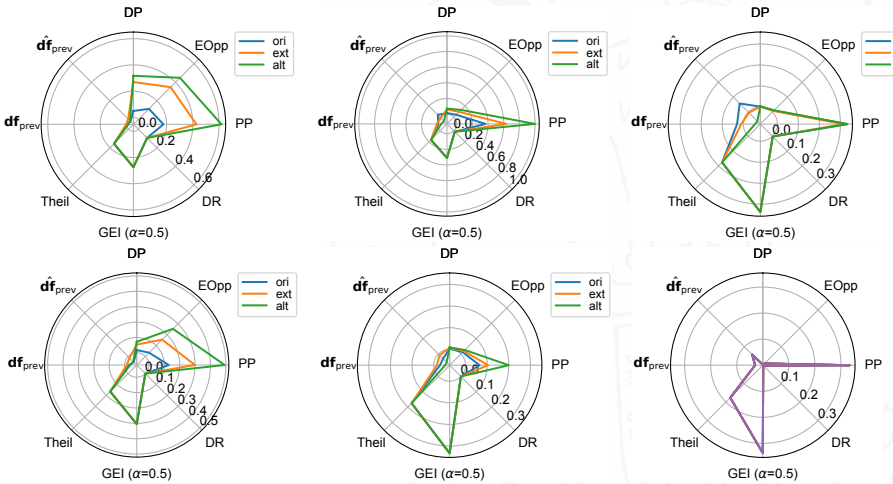


Figure 1: Comparison of fairness measures between their original definisions and their corresponding extension forms, on the Compas PPVR dataset. (a–e) Using bagging, AdaBoost, LightGBM, AdaFair (trained using #1 sen-att), and AdaFair (trained using #2 sen-att), respectively.

Traversal-based generalisation incurs computational burdens

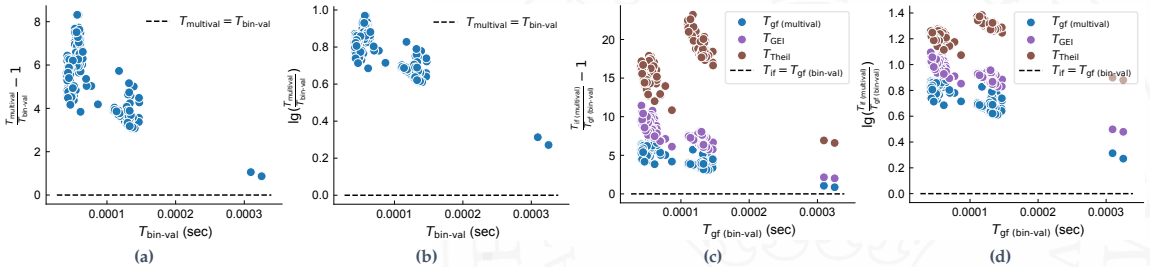


Figure 2: Time cost comparison of three commonly used group fairness measures and their extension forms, on Income, Compas PPR, and Compas PPVR datasets.
(a–b) Time cost comparison at different scales, and note that this is only for one 5- or 6-valued SA. Obviously, degenerating intersectional attributes ($\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 = \mathbb{Z}^{n_{a_1}} \times \mathbb{Z}^{n_{a_2}}$ where $n_{a_1}, n_{a_2} \geqslant 2$) into one "super" discrete SA through preprocessing is not an efficient way: It may be practical when both n_{a_1} and n_{a_2} are small enough, yet the computational cost increases exponentially as these values grow (e.g., if $n_{a_1} = 2$ and n_{a_2} changes from 2 to 6, $\mathcal{A}'$ transitions from $\mathbb{Z}^4$ to $\mathbb{Z}^{12}$).
(c–d) Time cost comparisons, including individual fairness measures that are suitable for one multi-valued SA, indicate that individual fairness has an even heavier computational burden.

Traversal-based generalisation incurs computational burdens

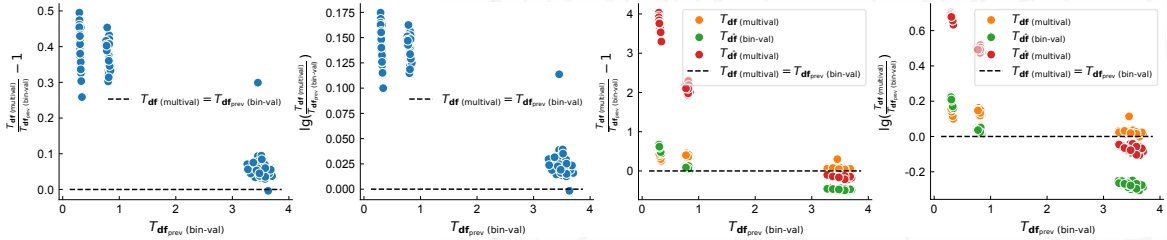


Figure 2: Time cost comparison of HFM for binary-value and multi-value cases, on Income, Compas PPR, and Compas PPVR datasets.
(a–b) Time cost comparisons of direct computation. (c–d) Comparisons including approximated results.[29]

[28] Bian and Luo, see n. 27; Bian, Luo, and Xu, see n. 27.
[29] Bian and Luo, see n. 27; Bian, Luo, and Xu, see n. 27.

Accuracy and fairness are not strictly incompatible

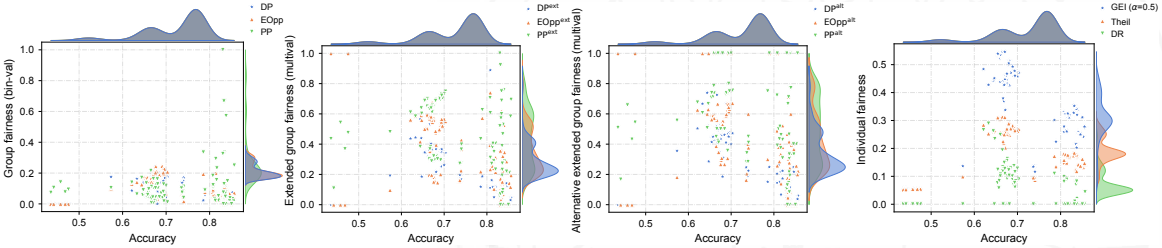


Figure 3: Scatter plot between performance (accuracy) and fairness. Note that on the *y*-axis, the smaller the better; on the *x*-axis, the larger the better. (a) Using three commonly used group fairness measures; (b) Using their first extension forms; (c) Using their second extension forms; (d) Using individual fairness measures.

Individual- and group- fairness are not inherently incompatible

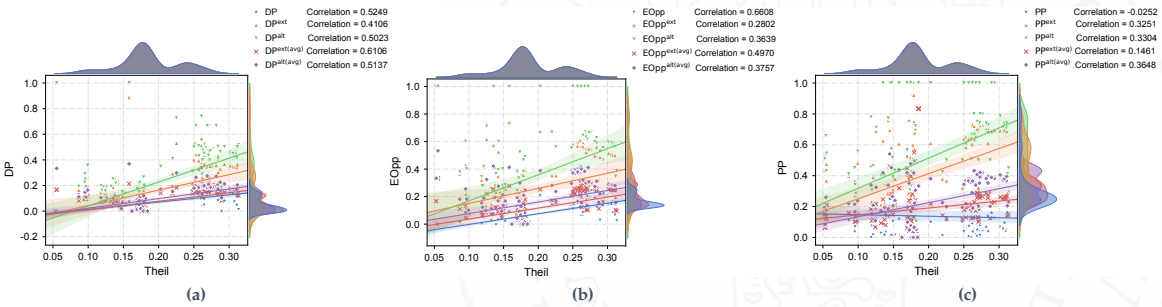


Figure 4: Relation between individual fairness and group fairness (DP, EOpp, and PP), on the Income, Compas PPR, and Compas PPVR datasets. Note that on both *x*- and *y*- axes, the smaller the better. (a–c) Using the Theil index as the individual fairness.

Individual- and group- fairness are not inherently incompatible

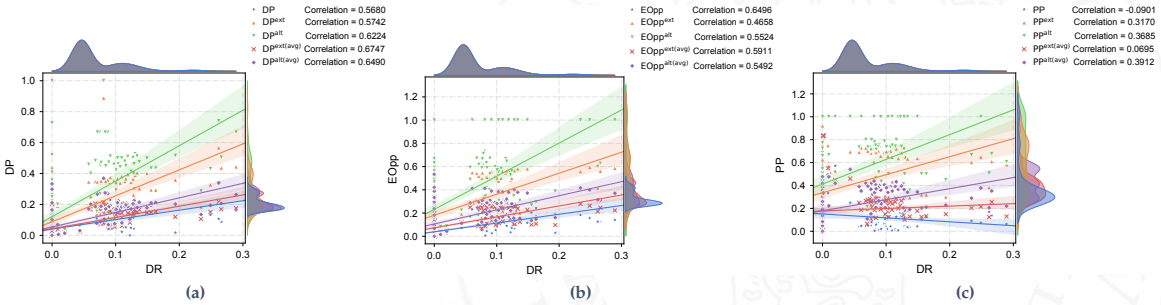


Figure 4: Relation between individual fairness and group fairness (DP, EOpp, and PP), on the Income, Compas PPR, and Compas PPVR datasets. Note that on both x- and y- axes, the smaller the better. (a–c) Using DR as the individual fairness.

Individual- and group- fairness are not inherently incompatible

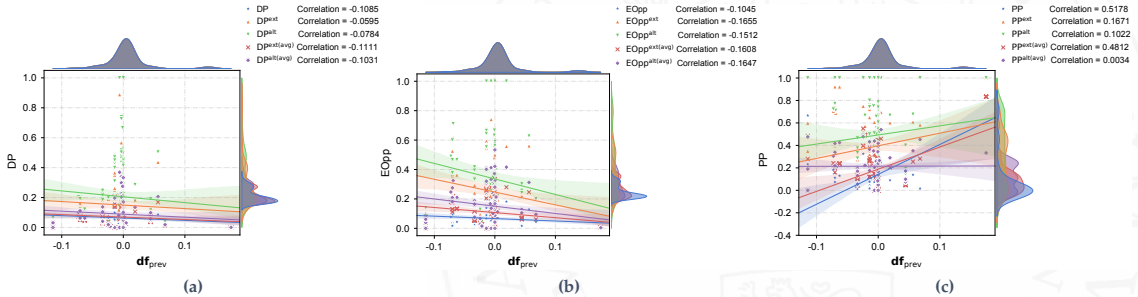


Figure 4: Relation between individual fairness and group fairness (DP, EOpp, and PP), on the Income, Compas PPR, and Compas PPVR datasets. Note that on both *x*- and *y*- axes, the smaller the better. (a–c) Using the previous HFM[31] as the individual fairness.

[30] Bian and Luo, see n. 27.
[31] Ibid.

Individual- and group- fairness are not inherently incompatible

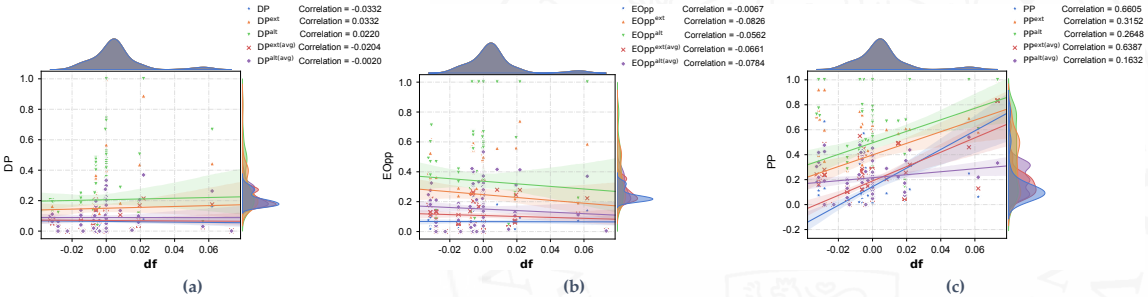


Figure 4: Relation between individual fairness and group fairness (DP, EOpp, and PP), on the Income, Compas PPR, and Compas PPVR datasets. Note that on both *x*- and *y*- axes, the smaller the better. (a–c) Using the maximum HFM[31] as the individual fairness.

[30] Bian, Luo, and Xu, see n. 27.
[31] Ibid.

Thanks! Questions?

Other distributive fairness

Fairness through unawareness[32]

A predictor is said to achieve *fairness through unawareness (FTU)* (or unconscious/unaware fairness) if all protected attributes $\mathcal{A}$ are excluded from the decision-making process.
Despite its compelling simplicity, this approach has a clear shortcoming: the remaining attributes $\mathcal{X}$ may contain discriminatory information analogous to $\mathcal{A}$ that may not be obvious at first, acting as *proxy* attributes. As a result, discrimination cannot be guaranteed to be eliminated.

Discriminative risk (DR)[33]

$$\mathrm{DR}(f) = \mathbb{E}[\mathbb{I}(f(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))], \tag{12}$$

where $\tilde{\boldsymbol{a}}$ is a perturbed $\boldsymbol{a}$, and $n_a \geqslant 1, |\mathcal{A}_i| \geqslant 2\,(i \in [n_a])$.

Harmonic fairness via manifold (HFM)[34] (three versions, see below)

[32] Dwork et al., see n. 10; Nina Grgić-Hlača et al. "The case for process fairness in learning: Feature selection for fair decision making". In: *NIPS symposium on machine learning and the law*. Vol. 1. 2. Barcelona, Spain. 2016, p. 11; Kusner et al., see n. 24; Gajane and Pechenizkiy, see n. 9.
[33] Bian and Zhang, see n. 26.
[34] Bian and Luo, see n. 27; Bian, Luo, and Xu, see n. 27.

Other distributive fairness

Harmonic fairness via manifold (HFM)[32]

Given a dataset $D = (\mathsf{X}, \mathsf{A}, \mathsf{Y})$, it has three versions: (1) the *previous* HFM for one bi-valued SA, and (2) the *maximal (resp. average)* HFM for several multi-valued SAs.

For one bi-valued SA $a_1 \in \mathcal{A}_1 = \{0, 1\}$, D is divided into $D_1 = \{(\check{\boldsymbol{x}}_a, y) \triangleq (\check{\boldsymbol{x}}, a_1, y) \in D \mid a_1 = 1\}$ and $\bar{D}_1 = D \setminus D_1$, then given a specific distance metric $\boldsymbol{d}(\cdot, \cdot)$ (e.g., the standard Euclidean metric), the previous HFM is

$$\mathbf{df}_{\text{prev}}(f) = \frac{\boldsymbol{g}_f(D_1, \bar{D}_1)}{\boldsymbol{g}(D_1, \bar{D}_1)} - 1\,, \tag{12}$$

where

$$\boldsymbol{g}_{\cdot}(D_1, \bar{D}_1; \ddot{y}) = \max\{\max_{(\check{\boldsymbol{x}}_a, y) \in D_1} \min_{(\check{\boldsymbol{x}}'_a, y') \in \bar{D}_1} \boldsymbol{d}((\check{\boldsymbol{x}}, \ddot{y}), (\check{\boldsymbol{x}}', \ddot{y}')) \max_{(\check{\boldsymbol{x}}'_a, y') \in \bar{D}_1} \min_{(\check{\boldsymbol{x}}_a, y) \in D_1} \boldsymbol{d}((\check{\boldsymbol{x}}, \ddot{y}), (\check{\boldsymbol{x}}', \ddot{y}'))\},$$

and $\boldsymbol{g}_f(D_1, \bar{D}_1) = \boldsymbol{g}_{\cdot}(D_1, \bar{D}_1; f(\check{\boldsymbol{x}}, a_1))$, $\boldsymbol{g}(D_1, \bar{D}_1) = \boldsymbol{g}_{\cdot}(D_1, \bar{D}_1; y)$ are two abbreviations for brevity.

[32] Bian and Luo, see n. 27; Bian, Luo, and Xu, see n. 27.

Other distributive fairness

Harmonic fairness via manifold (HFM) (cont.)

For one or more multi-valued SAs $\boldsymbol{a} \in \mathcal{A}$ where $n_a \geqslant 1$ and $|\mathcal{A}_i| \geqslant 2\,(i \in [n_a])$, the maximal (resp. average) HFM are

$$\mathbf{df} = \log\left(\frac{g_{f,\boldsymbol{a}}(D)}{g_{\boldsymbol{a}}(D)}\right), \tag{12a}$$

$$\mathbf{df}^{\text{avg}}(f) = \log\left(\frac{g^{\text{avg}}_{f,\boldsymbol{a}}(D)}{g^{\text{avg}}_{\boldsymbol{a}}(D)}\right), \tag{12b}$$

where

$$g_{\cdot,\boldsymbol{a}}(D;\ddot{\boldsymbol{y}}) = \max_{1\leqslant i\leqslant n_a} g_{\cdot,\boldsymbol{a}}(D,a_i;\ddot{\boldsymbol{y}}), \tag{13a}$$

$$g^{\text{avg}}_{\cdot,\boldsymbol{a}}(D;\ddot{\boldsymbol{y}}) = \tfrac{1}{n_a}\textstyle\sum_{i=1}^{n_a} g^{\text{avg}}_{\cdot,\boldsymbol{a}}(D,a_i;\ddot{\boldsymbol{y}}), \tag{13b}$$

$$g_{\cdot,\boldsymbol{a}}(D,a_i;\ddot{\boldsymbol{y}}) = \max_{i\in[n_{a_i}]}\{\max_{(\check{\boldsymbol{x}}_a,y)\in D_j}\min_{(\check{\boldsymbol{x}}'_a,y')\in\bar{D}_j} \boldsymbol{d}((\check{\boldsymbol{x}},\ddot{\boldsymbol{y}}),(\check{\boldsymbol{x}}',\ddot{\boldsymbol{y}}'))\},$$

$$g^{\text{avg}}_{\cdot,\boldsymbol{a}}(D,a_i;\ddot{\boldsymbol{y}}) = \tfrac{1}{n}\textstyle\sum_{j\in[n_{a_i}]}\sum_{(\check{\boldsymbol{x}},y)\in D_j}\min_{(\check{\boldsymbol{x}}',y')\in\bar{D}_j} \boldsymbol{d}((\check{\boldsymbol{x}},\ddot{\boldsymbol{y}}),(\check{\boldsymbol{x}}',\ddot{\boldsymbol{y}}')).$$

Note that $D_j = \{(\check{\boldsymbol{x}}_a, y) \in D | a_i = j\}$, $\bar{D}_j = D \backslash D_j$, and special case $g_{\cdot,\boldsymbol{a}}(D,a_i;\ddot{\boldsymbol{y}}) = g_{\cdot}(D_1,\bar{D}_1;\ddot{\boldsymbol{y}})$ when $\mathcal{A}_i = \{0,1\}$.

Procedural fairness

Here we use D to denote the set of all instances/members (or queried users of society), and S the set of all possible features that might be used in the decision-making process (in other words, $|S| \leqslant n_d + n_a$). Given a set of features S', let $f_{S'}$ denote the classifier that uses those features S'.

Feature-apriori fairness. For a given feature $s \in S$, let $D_s \subseteq D$ denote the set of all members that consider the feature s fair to use *without a priori knowledge* of how its usage affects outcomes. Then

$$\mathrm{PF}_{\mathrm{apr}}(f_{S'}) \triangleq \frac{|\bigcap_{s_i \in S'} D_{s_i}|}{|D|} . \tag{14}$$

Feature-accuracy fairness. (see below)
Feature-disparity fairness. (see below)

These three measures[32] accommodate scenarios with multiple SAs, each potentially having multiple values. Despite this advantage, they rely heavily on features and on a set of members/users who perceive these features as fair, which may still introduce hidden discrimination or human prejudice. Moreover, their computation is complex and time-consuming, as user judgments may evolve with learning, limiting their practical applicability.

[32] Grgić-Hlača et al., "The case for process fairness in learning: Feature selection for fair decision making", see n. 32; Nina Grgić-Hlača et al. "Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning". In: *AAAI*. vol. 32. 1. 2018.

Procedural fairness

Feature-accuracy fairness. Let $D_s^{\text{acc}} \subseteq D$ denote the set of all members that consider the feature s fair to be use *if it increases the accuracy of the classifier*. Note that typically $D_s \subseteq D_s^{\text{acc}}$ is expected, though this need not always hold exactly (due to either noise in estimating member preferences, or some members attaching some sort of negative connotation to the notion of accuracy). Then

$$\text{PF}_{\text{acc}}(f_{S'}) \triangleq \frac{|\bigcap_{s_i \in S'} \text{Cond}(D_{s_i}, D_{s_i}^{\text{acc}})|}{|D|}, \tag{14}$$

where

$$\text{Cond}(D_{s_i}, D_{s_i}^{\text{acc}}) = \begin{cases} D_{s_i}, & \text{if } \text{acc}(f_{S'}) \leqslant \text{acc}(f_{S' \setminus \{s_i\}}); \\ D_{s_i} \bigcup D_{s_i}^{\text{acc}} = D_{s_i}^{\text{acc}}, & \text{otherwise}. \end{cases}$$

Feature-disparity fairness. (cf. Appendices) Let $D_s^{\text{disp}} \subseteq D$ denote the set of all members that consider the feature s fair to use *even if it increases a measure of disparity* (i.e., disparate impact or disparate mistreatment) of the classifier. Typically $D_s^{\text{disp}} \subseteq D_s$ is expected, though this need not always hold strictly due to estimation error or other reasons. Let $\text{disp}(f_{S'})$ denote the disparity it induces, and then

$$\text{PF}_{\text{disp}}(f_{S'}) \triangleq \frac{|\bigcap_{s_i \in S'} \text{Cond}(D_{s_i}, D_{s_i}^{\text{disp}})|}{|D|}, \tag{15}$$

where

$$\text{Cond}(D_{s_i}, D_{s_i}^{\text{disp}}) = \begin{cases} D_{s_i}^{\text{disp}}, & \text{if } \text{disp}(f_{S'}) > \text{disp}(f_{S' \setminus \{s_i\}}); \\ D_{s_i} \bigcup D_{s_i}^{\text{disp}} = D_{s_i}, & \text{otherwise}. \end{cases}$$

Procedural fairness

Additionally,[32] proposed an FAE-based (feature attribution explanation) metric to assess group procedural fairness, which depends on the specific FAE techniques employed.

FAE-based group procedural fairness. A given dataset D is divided into two subsets by the values of a single SA, that is, $D_1 = \{(\mathbf{x}_i, a_{1i}, y_i) \in D \mid a_{1i} = 1\}$ and $D_2 = \{(\mathbf{x}_i, a_{1i}, y_i) \in D \mid a_{1i} = 0\}$. A local FAE function $g(\cdot)$ takes a model $f(\cdot)$ and an explained data point $(\mathbf{x}_i, a_{1i})$ as inputs and returns explanations (i.e., feature importance scores) $\mathbf{e}_i = g(f, \mathbf{x}_i, a_{1i}) \in \mathbb{R}^{n_d+1}$, where its j-th component e_{ij} is the importance score of the feature x_{ij} for the model's prediction $f(\mathbf{x}_i, a_{1i})$. For a distance measure $d_e(\cdot, \cdot)$ between two sets of FAE explanation results E_1 and E_2, then

$$\begin{aligned} \mathrm{GPF}_{\mathrm{FAE}} &= d_e(E_1, E_2)\,; \\ E_1 &= \{\mathbf{e}_i \mid \mathbf{e}_i = g(f, \mathbf{x}_i, a_{1i}), (\mathbf{x}_i, a_{1i}) \in D_1'\}\,, \\ E_2 &= \{\mathbf{e}_j \mid \mathbf{e}_j = g(f, \mathbf{x}_j, a_{1j}), (\mathbf{x}_j, a_{1j}) \in D_2'\}\,, \end{aligned} \tag{14}$$

where D_1' and D_2' are sets of n data points from D_1 and D_2, respectively, generated by [its Algorithm 1].

[32] Ziming Wang, Changwu Huang, and Xin Yao. "Procedural fairness in machine learning". In: *arXiv preprint arXiv:2404.01877* (2024).