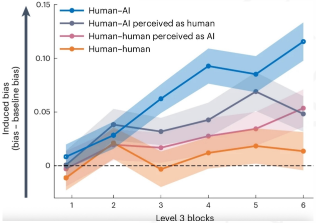
Why we need new fairness metrics and how to use them

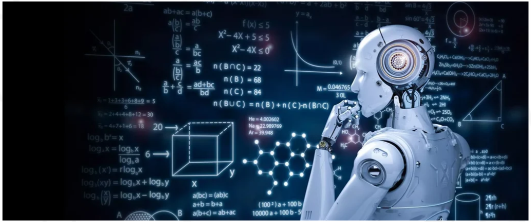
Yijun Bian

CopeNLU, NLP Section
Department of Computer Science
University of Copenhagen

12 December 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.

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[4] are:

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

These criteria are rarely satisfied all at once, except in degenerate cases.[5]

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

[5] 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.

Brief summary of existing fairness metrics

Table 1. Summary of existing fairness measures.[6,7]

Name of *measure*	Fairness type	**Meaning**		**Applicable situation(s)** in definition			**Non-binary handling**	
		quantitative	fairer	#label (n_c)	#sen-att (n_a)	#values per $\mathcal{A}_i$	$n_{a_i} > 2$	$n_a > 1$
Demographic parity (DP, 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 (EO)	*, group-	yes	lower value	binary	singular	bi-valued	yes	indirectly
Equality of opportunity (EOpp)	*, 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 (PP)	*, 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 [7]	*,[5]	yes	lower value	bi-/multi-	allows plural	multi-valued	—	—
Harmonic fairness via manifold [8]	*,[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

[6] Yijun Bian et al. "Algorithmic fairness: Not a purely technical but socio-technical property". In: *arXiv preprint arXiv: 2506.12556* (2025).

[7] Mark* indicates it belongs to *distributive* fairness; These two can measure discrimination from both group and individual fairness aspects.

Discriminative risk (DR)[9]

—from an individual-level aspect

Following the principle of individual fairness *(the treatment/evaluation on one instance should not change solely due to minor changes in its sensitive attributes)*, with an instance denoted by $\boldsymbol{x} = (\breve{\boldsymbol{x}}, \boldsymbol{a})$, the fairness quality of one hypothesis[8] $f(\cdot)$ could be evaluated by

$$\ell_{\text{bias}}(f, \boldsymbol{x}) = \mathbb{I}\big(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}) \big) \tag{1}$$

the indicator function

model prediction on the raw instance

model prediction when only sensitive attribute(s) are changed

non-sensitive attributes

sensitive attribute(s)

sensitive attribute(s) that are slightly perturbed
(the privileged $\leftrightarrow$ any one of the unprivileged)

similarly to the 0/1 loss, where $\tilde{\boldsymbol{a}}$ is a perturbed $\boldsymbol{a} = [a_1, ..., a_{n_a}]^{\mathsf{T}}$, $a_i \in \mathcal{A}_i$, $n_a \geqslant 1$, and $|\mathcal{A}_i| \geqslant 2$. Note that Eq. (1) is evaluated on only one instance with sensitive attributes $\boldsymbol{x}$.

[8] The hypothesis used in this equation could indicate an individual classifier or an ensemble classifier.

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

Discriminative risk (DR)[9]

—from an individual-level aspect
—from a group-level aspect

To describe this characteristic of the hypothesis on multiple instances, then the empirical discriminative risk on one dataset S is expressed as $\hat{\mathcal{L}}_{\text{bias}}(f,S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f,\boldsymbol{x}_i)$, and the true discriminative risk[8] of the hypothesis over a data distribution is $\mathcal{L}_{\text{bias}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell_{\text{bias}}(f,\boldsymbol{x})]$, respectively. Note that the empirical DR on S is an unbiassed estimation of the true DR.

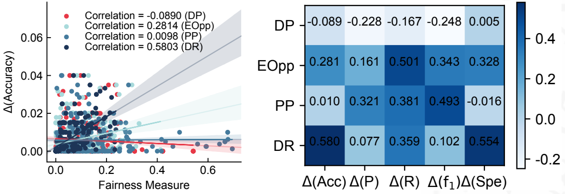
$$\ell_{\text{bias}}(f,\boldsymbol{x}) = \mathbb{I}(f(\breve{\boldsymbol{x}},\boldsymbol{a}) \neq f(\breve{\boldsymbol{x}},\tilde{\boldsymbol{a}}))$$
$$\mathrm{DR}(f) = \mathbb{E}[\mathbb{I}(f(\breve{\boldsymbol{x}},\boldsymbol{a}) \neq f(\breve{\boldsymbol{x}},\tilde{\boldsymbol{a}}))]$$

- Widely applicable, allowing one or more SAs, and each SA allowing binary or multiple values
- Different from existing (group/individual/counterfactual) fairness measures
- Limitations: small values; computational results may be affected somehow by a randomness factor

[8]The instances from S are independent identically distributed (i.i.d.) drawn from an input/feature-output/label space $\mathcal{X}\times\mathcal{Y}$ according to an unknown distribution $\mathcal{D}$.

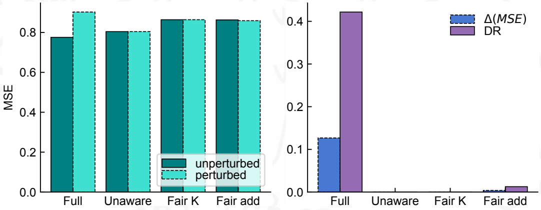
[9]Bian and Zhang, see n. 9.

Validating *DR*, a fairness quality measure



(a) **(b)**

Figure 1. Comparison of the proposed discriminative risk (DR) with three group fairness measures, that is, DP, EOpp, and PP.
(a) Scatter diagrams with the degree of correlation, where the *x*- and *y*-axes are different fairness measures and the variation of accuracy between the raw and disturbed data. (b) Correlation among multiple criteria. Note that correlation here is calculated based on the results from all datasets.

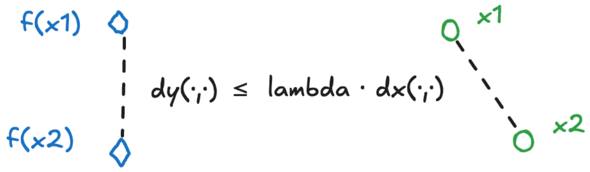


(a) **(b)**

Figure 2. Example: law school success.
(a) Test MSE of different models, where 'unperturbed' and 'perturbed' denote the results obtained from the original and disturbed data respectively. (b) The comparison between the change in MSE and *DR*, which suggests that DR ≈ 0 when the corresponding model satisfies or nearly satisfies counterfactual fairness.

Differences from existing fairness measures

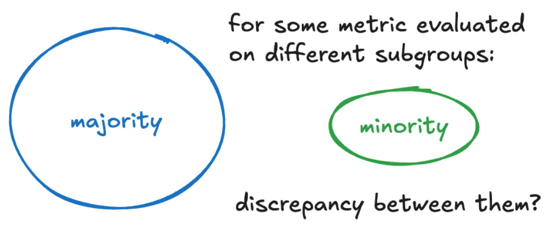
- Two distinctions from *individual fairness* measures
  1. relies on the choice of similarity/distance metric
  2. instance pairs in comparison coming from original data



- Two distinctions from *group fairness* measures
- Four distinctions from *causal fairness*

Differences from existing fairness measures

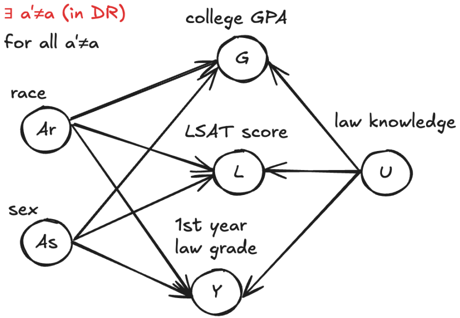
- Two distinctions from *individual fairness* measures
- Two distinctions from *group fairness* measures
  1. works for only one sensitive attribute (usually bi-valued)
  2. computing separately for each subgroup, then difference



- Four distinctions from *causal fairness*

Differences from existing fairness measures

- Two distinctions from *individual fairness* measures
- Two distinctions from *group fairness* measures
- Four distinctions from *causal fairness*
  1. works for only one sensitive attribute (although possibly multi-valued)
  2. based on causal models/graphs, not a quantitative measure
  3. non-sensitive attributes may vary with it in counterfactual fairness
  4. conditions for achieving them are stronger



Differences from existing fairness measures

- Two distinctions from *individual fairness* measures
- Two distinctions from *group fairness* measures
- Four distinctions from *causal fairness*

$$\ell_{\text{bias}}(f,\boldsymbol{x}) = \mathbb{I}(f(\breve{\boldsymbol{x}},\boldsymbol{a}) \neq f(\breve{\boldsymbol{x}},\tilde{\boldsymbol{a}}))$$

$$\hat{\mathcal{L}}_{\text{bias}}(f,S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f,\boldsymbol{x}_i)$$

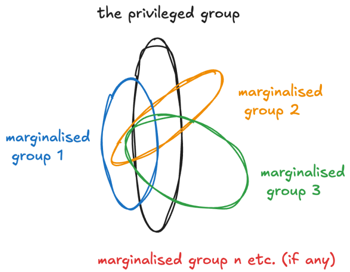
$$\mathcal{L}_{\text{bias}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\,\ell_{\text{bias}}(f,\boldsymbol{x})\,]$$

$$\mathcal{L}'_{\text{bias}}(f) = |\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}|\boldsymbol{a}=1}[\ell_{\text{bias}}(f,\boldsymbol{x})] - \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}|\boldsymbol{a}=0}[\ell_{\text{bias}}(f,\boldsymbol{x})]|$$

- Similarities that *DR* shares with the existing fairness measures
  - follows the same principle as *individual fairness* measures
  - is computed over a group of instances (like one dataset or a data distribution)
  - indicates the discrimination level from a statistical/demographic perspective

Harmonic fairness via manifolds (HFM) [11]

If we view the *instances (with the same value of sensitive attributes)* as *data points on certain manifold(s)*, the manifold representing members from the marginalised/unprivileged group(s) is supposed to be as close as possible to that representing members from the privileged group.

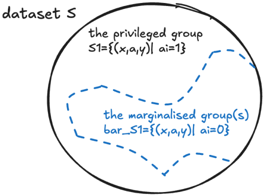


Given a dataset $S = (\mathsf{X}, \mathsf{A}, \mathsf{Y})$, we measure fairness with respect to the sensitive attribute(s) (SAs) called HFM[10] in three variants: (i) the *previous* HFM for a single binary SA; and (ii–iii) the *maximal (resp. average)* HFM over multiple (possibly multi-valued) SAs. HFM is built upon the concept of distances between sets, where we use the Hausdorff distance, recorded as **D**, to *evaluate the discrepancy among groups divided by sensitive attributes*.

[10] indicating difference from both individual- and group- apsects

[11] 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); 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).

The previous HFM —for one *bi-valued* SA



For one bi-valued SA $a_1 \in \mathcal{A}_1 = \{0,1\}$, S is divided into $S_1 = \{(x,y) \triangleq (\breve{x}, a_1, y) \in D \mid a_1 = 1\}$ and $\bar{S}_1 = S \setminus S_1$, then given a specific distance metric $d(\cdot,\cdot)$[12] on the feature space, the previous HFM is

$$\mathbf{df}_{\text{prev}}(f) = \mathbf{D}_f(S_1, \bar{S}_1) / \mathbf{D}(S_1, \bar{S}_1) - 1\,, \tag{2}$$

where

the privileged group (S_1); the marginalised/unprivileged group(s) ($\bar{S}_1$); works for both the true label y and the prediction $\hat{y}$ of a trained classifier $f(\cdot)$ ($\ddot{y}$)

$$\mathbf{D}.(S_1, \bar{S}_1; \ddot{y}) \triangleq \max\Big\{ \max_{(x,y)\in S_1} \underbrace{\min_{(x',y')\in \bar{S}_1} \mathbf{d}\big((\breve{x},\ddot{y}),(\breve{x}',\ddot{y}')\big)}_{\text{to find the nearest data point in } \bar{S}_1},\ \max_{(x',y')\in \bar{S}_1} \min_{(x,y)\in S_1} \mathbf{d}\big((\breve{x},\ddot{y}),(\breve{x}',\ddot{y}')\big) \Big\}\,, \tag{3}$$

is an approximation of the distance between the manifold of unprivileged groups and that of the privileged group, and $\mathbf{D}_f(S_1, \bar{S}_1) = \mathbf{D}.(S_1, \bar{S}_1; f(\breve{x}, a_1))$, $\mathbf{D}(S_1, \bar{S}_1) = \mathbf{D}.(S_1, \bar{S}_1; y)$ are two abbreviations for brevity.

[12] Here we use the standard Euclidean metric. In fact, any two metrics $\mathbf{d}_1, \mathbf{d}_2$ derived from norms on the Euclidean space $\mathbb{R}^d$ are equivalent in the sense that there are positive constants c_1, c_2 such that $c_1\mathbf{d}_1(x,y) \leqslant \mathbf{d}_2(x,y) \leqslant c_2\mathbf{d}_1(x,y)$ for all $x, y \in \mathbb{R}^d$.

HFM —over multiple (possibly multi-valued) SAs

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

$$\mathbf{df} = \log\left(\mathbf{D}_{f,\boldsymbol{a}}(S) / \mathbf{D}_{\boldsymbol{a}}(S)\right) , \tag{4a}$$

$$\mathbf{df}^{\text{avg}}(f) = \log\left(\mathbf{D}_{f,\boldsymbol{a}}^{\text{avg}}(S) / \mathbf{D}_{\boldsymbol{a}}^{\text{avg}}(S)\right) , \tag{4b}$$

where

$$\mathbf{D}_{\cdot,\boldsymbol{a}}(S;\ddot{y}) = \max\nolimits_{1 \leqslant i \leqslant n_a} \mathbf{D}_{\cdot,\boldsymbol{a}}(S, a_i; \ddot{y}) , \tag{5a}$$

$$\mathbf{D}_{\cdot,\boldsymbol{a}}^{\text{avg}}(S;\ddot{y}) = \tfrac{1}{n_a} \sum\nolimits_{i=1}^{n_a} \mathbf{D}_{\cdot,\boldsymbol{a}}^{\text{avg}}(S, a_i; \ddot{y}) , \tag{5b}$$

$$\mathbf{D}_{\cdot,\boldsymbol{a}}(S, a_i; \ddot{y}) = \max\nolimits_{j \in [n_{a_i}]} \{ \max\nolimits_{(\boldsymbol{x},y) \in S_j} \overbrace{\min\nolimits_{(\boldsymbol{x}',y') \in \bar{S}_j} \boldsymbol{d}((\breve{\boldsymbol{x}},\ddot{y}),(\breve{\boldsymbol{x}}',\ddot{y}'))}^{\text{to fine the nearest data point in } \bar{S}_j} \},$$

$$\mathbf{D}_{\cdot,\boldsymbol{a}}^{\text{avg}}(S, a_i; \ddot{y}) = \tfrac{1}{n} \sum\nolimits_{j \in [n_{a_i}]} \sum\nolimits_{(\breve{\boldsymbol{x}},y) \in S_j} \min\nolimits_{(\breve{\boldsymbol{x}}',y') \in \bar{S}_j} \boldsymbol{d}((\breve{\boldsymbol{x}},\ddot{y}),(\breve{\boldsymbol{x}}',\ddot{y}')) .$$

Note that $S_j = \{(\boldsymbol{x},y) \in S | a_i = j\}$, $\bar{S}_j = S \setminus S_j$, and special case $\mathbf{D}_{\cdot,\boldsymbol{a}}(S, a_i; \ddot{y}) = \mathbf{D}_{\cdot}(S_1, \bar{S}_1; \ddot{y})$ when $\mathcal{A}_i = \{0, 1\}$.

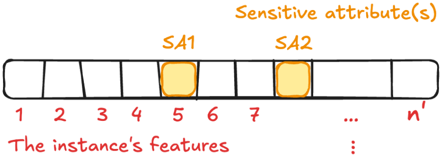
Interim summary[13]

RQ 1. *How to properly measure the discriminative level of a classifier from both individual and group fairness aspects?*
RQ 2. *How to efficiently measure the extra discrimination introduced in learning by a classifier?*

	Work 1	**Work 2**	
	previous	maximal	average
Distance between sets	$\mathbf{D}(S_1, \bar{S}_1; \ddot{y})$	$\mathbf{D}_{\cdot,\boldsymbol{a}}(S, a_i)$	$\mathbf{D}^{\text{avg}}_{\cdot,\boldsymbol{a}}(S, a_i)$
		$\mathbf{D}_{\cdot,\boldsymbol{a}}(S)$	$\mathbf{D}^{\text{avg}}_{\cdot,\boldsymbol{a}}(S)$
HFM (fairness measure)	$\mathbf{df}_{\text{prev}}(f)$	$\mathbf{df}(f)$	$\mathbf{df}^{\text{avg}}(f)$
Approximation for one SA	*AcceleDist*		*AcceleDist*
	ApproxDist		*ApproxDist*
Approximation for several SA			*ExtendDist*

[13] Bian and Luo, see n. 11; Bian, Luo, and Xu, see n. 11.

ExtendDist[14] [Work 2 in the series of HFM]



Algorithm 1. Approximation of extended distance between sets for several sensitive attributes with multiple values, aka. $\mathtt{ExtendDist}(\{(\breve{x}_i, \boldsymbol{a}_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n; m_1, m_2)$,

Input: Dataset $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n = \{(\breve{x}_i, \boldsymbol{a}_i, y_i)\}_{i=1}^n$ where $\boldsymbol{a}_i = [a_{i,1}, a_{i,2}, ..., a_{i,n_a}]^\mathsf{T}$, prediction of S by the classifier $f(\cdot)$ that has been trained, that is, $\{\hat{y}_i\}_{i=1}^n$, and two hyper-parameters m_1 and m_2 as the designated numbers for repetition and comparison respectively

Output: Approximation of $\mathbf{D}_{\cdot,\boldsymbol{a}}(S)$ and $\mathbf{D}_{\cdot,\boldsymbol{a}}^{\text{avg}}(S)$

1: **for** j from 1 to n_a **do**

2: $\quad d_{\max}^{(j)}, d_{\text{avg}}^{(j)} = \mathtt{ApproxDist}(\{(\breve{x}_i, a_{i,j})\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n; m_1, m_2)$

3: **return** $\max_{1 \leqslant j \leqslant n_a}\{d_{\max}^{(j)} \mid j \in [n_a]\}$ and $\frac{1}{n_a}\sum_{j=1}^{n_a} d_{\text{avg}}^{(j)}$

[14] Bian, Luo, and Xu, see n. 11.

ApproxDist[15] [Work 1&2 in the series of HFM]

Algorithm 2. (Simplified) Approximation of distance between sets, aka. $\texttt{ApproxDist}\left(\{(\breve{x}_i, a_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n; m_1, m_2\right)$

Input: Dataset $S = \{(x_i, y_i)\}_{i=1}^n = \{(\breve{x}_i, a_i, y_i)\}_{i=1}^n$, prediction of S by the classifier $f(\cdot)$ that has been trained, that is, $\{\hat{y}_i\}_{i=1}^n$, and two hyper-parameters m_1 and m_2 as the designated numbers for repetition and comparison respectively

Output: Approximation of distance $\mathbf{D}_\cdot(S_1, \bar{S}_1)$ in Eq. (3)

1: **for** j from 1 to m_1 **do**

2: Take a random vector $\boldsymbol{w}$ from the space $\mathcal{W} = \{\boldsymbol{w} = [w_0, w_1, ..., w_{n_x}]^\mathsf{T} \mid \sum_{i=0}^{n_x} |w_i| = 1\} \subseteq [-1, 1]^{1+n_x}$

3: $d_{\max}^j = \texttt{AcceleDist}\left(\{(\breve{x}_i, a_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n, \boldsymbol{w}; m_2\right)$

4: **return** $\min\{d_{\max}^j \mid j \in [m_1]\}$

Algorithm 2. Approximation of distance between sets (for one sensitive attribute with multiple values), aka. $\texttt{ApproxDist}\left(\{(\breve{x}_i, a_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n; m_1, m_2\right)$

Input: Dataset $S = \{(x_i, y_i)\}_{i=1}^n = \{(\breve{x}_i, \boldsymbol{a}_i, y_i)\}_{i=1}^n$, prediction of S by the classifier $f(\cdot)$ that has been trained, that is, $\{\hat{y}_i\}_{i=1}^n$, and two hyper-parameters m_1 and m_2 as the designated numbers for repetition and comparison respectively

Output: Approximation of $\mathbf{D}_{\cdot,\boldsymbol{a}}(S, a_i)$ and $\mathbf{D}_{\cdot,\boldsymbol{a}}^{\text{avg}}(S, a_i)$

1: **for** j from 1 to m_1 **do**

2: Take two orthogonal vectors $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$ where each $\boldsymbol{w}_k \in [-1, +1]^{1+n_x}$ $(k = \{0, 1\})$

3: **for** k from 0 to 1 **do**

4: $t_{\max}^k, t_{\text{avg}}^k = \texttt{AcceleDist}\left(\{(\breve{x}_i, a_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n, \boldsymbol{w}_k; m_2\right)$

5: $d_{\max}^j = \min\{t_{\max}^k \mid k \in \{0, 1\}\} = \min\{t_{\max}^0, t_{\max}^1\}$

6: $d_{\text{avg}}^j = \min\{t_{\text{avg}}^k \mid k \in \{0, 1\}\} = \min\{t_{\text{avg}}^0, t_{\text{avg}}^1\}$

7: **return** $\min\{d_{\max}^j \mid j \in [m_1]\}$ and $\frac{1}{n}\min\{d_{\text{avg}}^j \mid j \in [m_1]\}$

[15] Bian and Luo, see n. 11; Bian, Luo, and Xu, see n. 11.

Distance *approximation* for Euclidean spaces

We observe that *the distance between similar data points tends to be closer than others after projecting them onto a general one-dimensional linear subspace* (refer to[16]).

To estimate the distance between data points inside $\mathcal{X} \times \mathcal{Y}$,

$$g(\boldsymbol{x}, \ddot{y}; \boldsymbol{w}) = g(\breve{\boldsymbol{x}}, \boldsymbol{a}, \ddot{y}; \boldsymbol{w}) = [\ddot{y}, x_1, \ldots, x_{n_x}]^\mathsf{T} \boldsymbol{w}\,, \tag{6}$$

where

- a random projection $g : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$
- a non-zero random vector $\boldsymbol{w} = [w_0, w_1, \ldots, w_{n_x}]^\mathsf{T}$

That is to say, after sorting all the projected data points on $\mathbb{R}$, it is likely that *for one instance* $(\boldsymbol{x}, y)$ *in* S_j*, the desired instance* $\operatorname{argmin}_{(\boldsymbol{x}',y') \in \bar{S}_j} \mathbf{d}\big((\breve{\boldsymbol{x}}, y), (\breve{\boldsymbol{x}}', y')\big)$ *would be somewhere near it after the projection, and vice versa*. Thus, searching for it could be accelerated by checking several adjacent instances rather than traversing the whole dataset.

[16] Bian and Luo, see n. 11, Lemma 1.

AcceleDist[17] [Work 1&2 in the series of HFM]

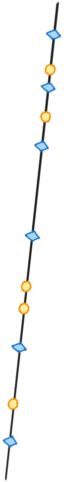
Algorithm 3. Acceleration sub-procedure in approximation, aka. `AcceleDist` $(\{(\breve{x}_i, a_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n, \boldsymbol{w}; m_2)$

Input: Data points $\{(\breve{x}_i, a_i)\}_{i=1}^n$, its corresponding value $\{\ddot{y}_i\}_{i=1}^n$, where $\ddot{y}_i$ could be its true label y_i or prediction $\hat{y}_i$ by the classifier $f(\cdot)$, a random vector $\boldsymbol{w}$ for projection, and a hyper-parameter m_2 as the designated number for comparison

Output: Approximation of distance $\mathbf{D}.(S_0, S_1)$ in Eq. (3)

Output: Approximation of $\mathbf{D}_{\cdot,a}(S, a_i)$ and $n\mathbf{D}_{\cdot,a}^{\text{avg}}(S, a_i)$

```
Project data points onto a one-dimensional space based on Eq. (6), in order to obtain {g(x_i, ÿ_i; w)}_{i=1}^n
Sort original data points based on {g(x_i, ÿ_i; w)}_{i=1}^n as their corresponding values, in ascending order
for i from 1 to n do
    Set the anchor data point (x_i, ÿ_i) in this round
    // If a_i = j (marked for clarity), in order to approximate min_{(x',y')∈S̄_j} d((x̆_i, ÿ_i), (x̆', ÿ'))
    Compute the distances d((x̆_i, ÿ_i), ·) for at most m_2 nearby data points that meets a ≠ a_i and g(x̆, ÿ; w) ⩽ g(x̆_i, ÿ_i; w)
    Find the minimum among them, recorded as d^s_min
    Compute the distances d((x̆_i, ÿ_i), ·) for at most m_2 nearby data points that meets a ≠ a_i and g(x, ÿ; w) ⩾ g(x_i, ÿ_i; w)
    Find the minimum among them, recorded as d^r_min
    d^(i)_min = min{d^s_min, d^r_min}
return max{d^(i)_min | i ∈ [n]}
return max{d^(i)_min | i ∈ [n]} and ∑_{i=1}^n d^(i)_min
```



[17] Bian and Luo, see n. 11; Bian, Luo, and Xu, see n. 11.

Binarisation underestimates discrimination

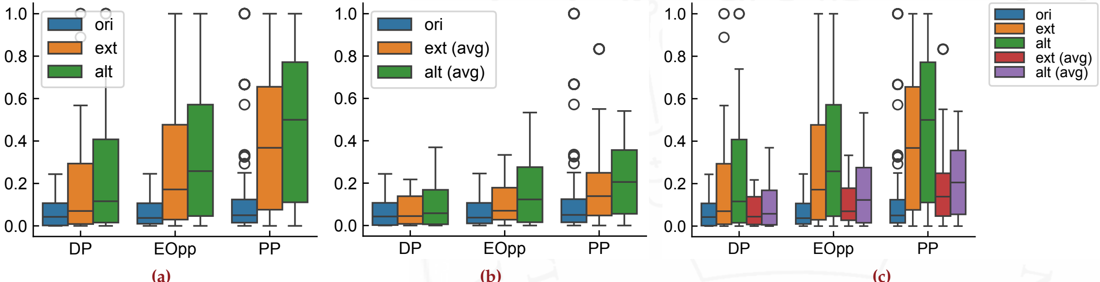


Figure 3. Comparison of three commonly used group fairness measures and their extensions.[18]

$$|\mathbb{P}(f(\breve{x},a_1)=1 \mid a_1=0)-\mathbb{P}(f(\breve{x},a_1)=1 \mid a_1=1)| \leqslant \varepsilon\,, \tag{7a}$$

$$|\mathbb{P}(f(\breve{x},a_1)=1 \mid a_1\neq 1)-\mathbb{P}(f(\breve{x},a_1)=1 \mid a_1=1)|\,, \tag{7b}$$

$$\max_{j\in\mathcal{A}_1} |\mathbb{P}(f(\breve{x},a_1)=1 \mid a_1=j)-\mathbb{P}(f(\breve{x},a_1)=1)|\,, \tag{7c}$$

$$\max_{j,k\in\mathcal{A}_1,j\neq k} |\mathbb{P}(f(\breve{x},a_1)=1 \mid a_1=j)-\mathbb{P}(f(\breve{x},a_1)=1 \mid a_1=k)|\,. \tag{7d}$$

[18] on Income, Compas PPR, and Compas PPVR datasets. (a–b) Comparison between binarisation and the two extension forms, analogously to Eq. (7c) and (7d); note that binarisation is equivalent to their original definitions like (7a). (c–d) Comparison between binarisation and their corresponding average forms. (e) Comparison between binarisation and all four extension formulas.

Binarisation underestimates discrimination

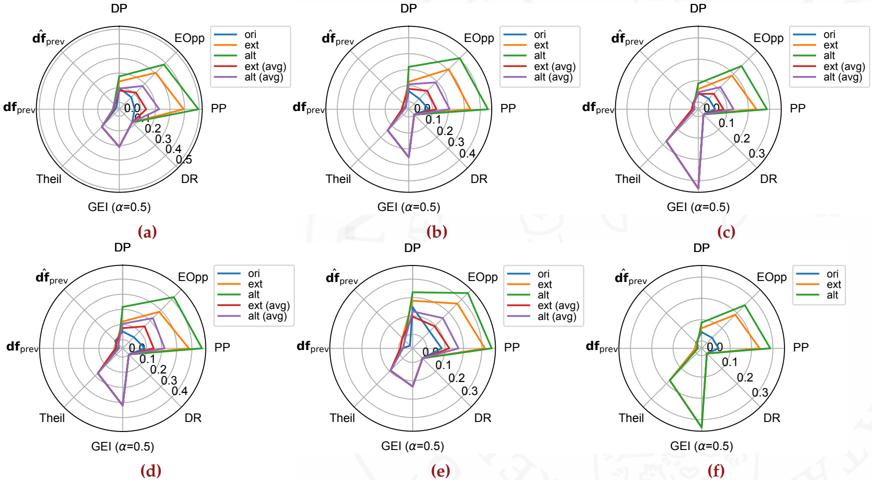
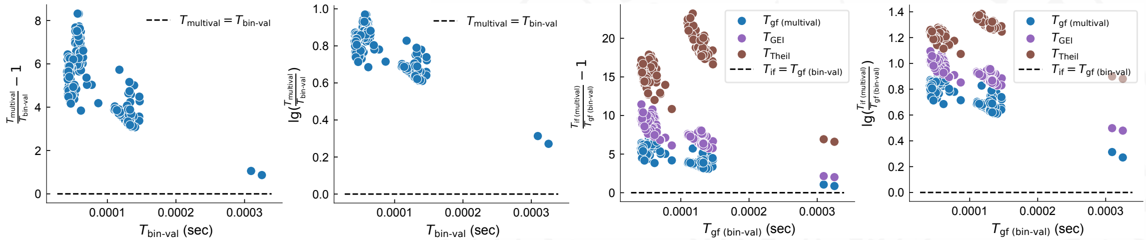


Figure 3. This pattern recurs across datasets regardless of the learning algorithms in use, and suggests that it oversimplifies the structure of disadvantage and can systematically underestimate discrimination.[18]

[18] on the Income dataset, using: (a–e) bagging, AdaBoost, LightGBM, AdaFair (trained using #1 sen-att), and AdaFair (trained using #2 sen-att), respectively; (f) LightGBM.

Traversal-based generalisation incurs computational burdens



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

Figure 4. Time cost comparison of three commonly used group fairness measures and their extension forms, at different scales, 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

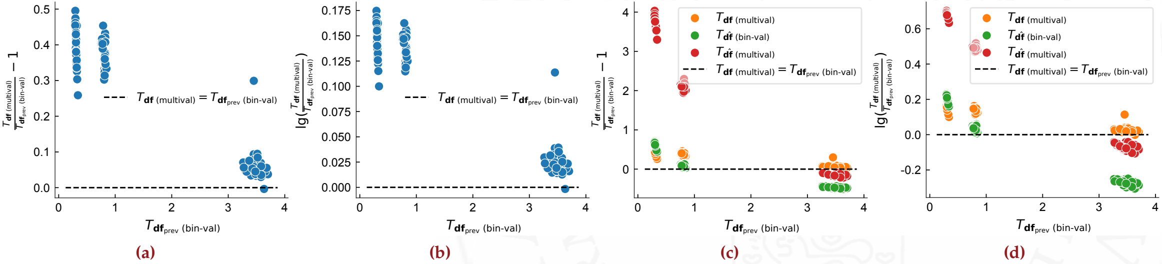


Figure 4. Time cost comparison of HFM for binary-value and multi-value cases, at different scales, on Income, Compas PPR, and Compas PPVR datasets.
(a–b) Time cost comparisons of direct computation. (c–d) Comparisons including approximated results.[19]

[19] Bian and Luo, see n. 11; Bian, Luo, and Xu, see n. 11.

Individual- and group- fairness are not inherently incompatible

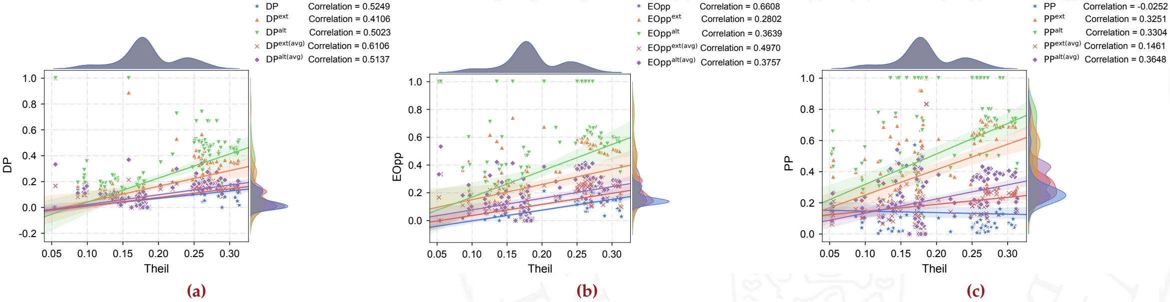


Figure 5. Relation between individual fairness and group fairness (DP, EOpp, and PP), on the Income, Compas PPR, and Compas PPVR datasets. Note that on both *x*- and *y*- axes, the smaller the better. (a–c) The individual fairness used here is the Theil index[20].

[20] Christian Haas. "The price of fairness - A framework to explore trade-offs in algorithmic fairness". In: *ICIS*. Association for Information Systems. 2019.

Individual- and group- fairness are not inherently incompatible

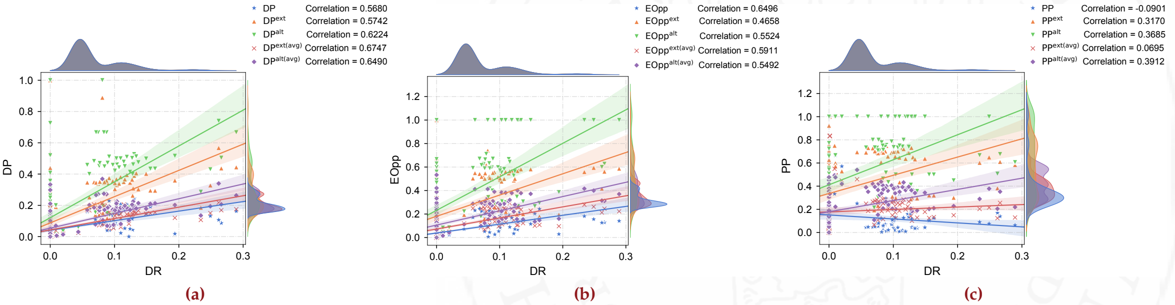


Figure 5. Relation between individual fairness and group fairness (DP, EOpp, and PP), on the Income, Compas PPR, and Compas PPVR datasets. Note that on both *x*- and *y*- axes, the smaller the better. (a–c) The individual fairness used here is discriminative risk (DR).[20]

[20]Bian and Zhang, see n. 9.

Overall framework

FairSHAP: Preprocessing for Fairness Through Attribution-Based Data Augmentation[21]

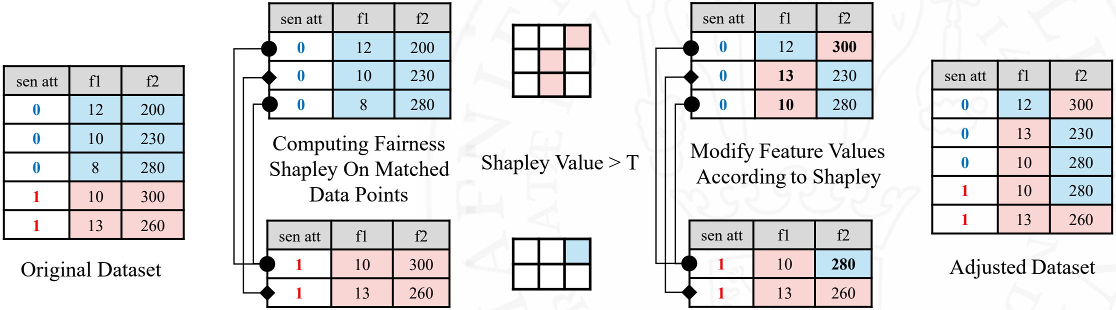


Figure 6. Overall framework of ***FairSHAP***: **(Left)** Training data are first split by sensitive attribute and aligned via nearest-neighbour matching to produce paired instances; **(Right)** For each target group, feature values whose Shapley value exceeds a threshold are adjusted to reduce DR, and the modified instances from both groups are recombined into an augmented, fairness-improved training set.

[21] Lin Zhu, Yijun Bian, and Lei You. "FairSHAP: Preprocessing for fairness through attribution-based data augmentation". In: *arXiv preprint arXiv:2505.11111* (2025). Under review.

Qualitative results

Table 2. Compare FairSHAP with other fairness mitigation methods across different datasets.[22]

Dataset (s.a.)	Methods	Accuracy	DR	DP	EO	PQP	Data Fidelity	TrainingAR	TestAN
German (sex)	Baseline	0.6650±0.0257	0.0785±0.0211	0.0512±0.0346	0.1287±0.0590	0.1341±0.0486	—	—	No
	CR	0.6680±0.0238	**0.0028±0.0029**	0.0844±0.0557	0.1559±0.0609	**0.0723±0.0330**	0.0183±0.0211	0.9615	Yes
	DIR	**0.6720±0.0337**	0.0966±0.0112	0.0946±0.0373	0.1737±0.0729	0.1529±0.0634	0.0155±0.0440	0.0774	Yes
	FairSHAP	0.6630±0.0275	0.0243±0.0112	**0.0301±0.0347**	**0.1126±0.0783**	0.1852±0.1074	**0.0049±0.0085**	**0.0156**	**No**
COMPAS (sex)	Baseline	**0.6698±0.0051**	0.0883±0.0064	0.1548±0.0241	0.1243±0.0510	0.0492±0.0084	—	—	No
	CR	0.6679±0.0045	**0.0082±0.0070**	0.1407±0.0248	0.1291±0.0317	0.0714±0.0517	0.0189±0.0193	0.9174	Yes
	DIR	0.6644±0.0098	0.1150±0.0091	**0.1155±0.0239**	**0.0952±0.0359**	0.0747±0.0370	0.0387±0.0640	0.0650	Yes
	FairSHAP	0.6609±0.0106	0.0629±0.0091	0.1326±0.0407	0.0985±0.0603	**0.0452±0.0383**	**0.0025±0.0048**	**0.0113**	**No**
COMPAS (race)	Baseline	**0.6689±0.0108**	0.0995±0.0076	0.1436±0.0209	0.1438±0.0233	0.0522±0.0406	—	—	No
	CR	0.6611±0.0112	**0.0418±0.0092**	0.1502±0.0341	0.1621±0.0530	0.0592±0.0367	0.0250±0.0222	0.892	Yes
	DIR	0.6149±0.0286	0.1185±0.0181	0.1359±0.1241	**0.1117±0.0945**	**0.0399±0.0338**	0.0512±0.0736	0.0701	Yes
	FairSHAP	0.6627±0.0069	0.0842±0.0049	**0.1344±0.0332**	0.1568±0.0343	0.0508±0.0469	**0.0040±0.0055**	**0.0126**	**No**
Adult (sex)	Baseline	**0.8722±0.0033**	0.0315±0.0037	0.1805±0.0066	0.0735±0.0275	**0.0275±0.0321**	—	—	No
	CR	0.8706±0.0029	**0.0000±0.0000**	0.1824±0.0055	0.0955±0.0243	0.0278±0.0173	0.0167±0.0391	0.9887	Yes
	DIR	0.8550±0.0067	0.0499±0.0076	0.1607±0.0157	0.0772±0.0624	0.0360±0.0253	0.0046±0.0417	0.0081	Yes
	FairSHAP	0.8692±0.0046	0.0273±0.0047	**0.1558±0.0130**	**0.0393±0.0254**	0.0474±0.0319	**0.0010±0.0073**	**0.0012**	**No**
Adult (race)	Baseline	**0.8721±0.0033**	0.0398±0.0025	0.1034±0.0110	0.0808±0.0326	0.0302±0.0265	—	—	No
	CR	0.8713±0.0033	**0.0000±0.0000**	0.1008±0.0115	0.0983±0.0389	0.0480±0.0235	0.0300±0.0450	0.962	Yes
	DIR	0.8320±0.0173	0.0740±0.0209	**0.0703±0.0515**	0.0871±0.0355	0.0482±0.0730	0.0252±0.0480	0.0089	Yes
	FairSHAP	0.8720±0.0023	0.0284±0.0017	0.0851±0.0155	**0.0287±0.0277**	**0.0259±0.0318**	**0.0030±0.0084**	**0.0014**	**No**

[22] CR: CorrelationRemover; DIR: DisparateImpactRemover.

Qualitative results

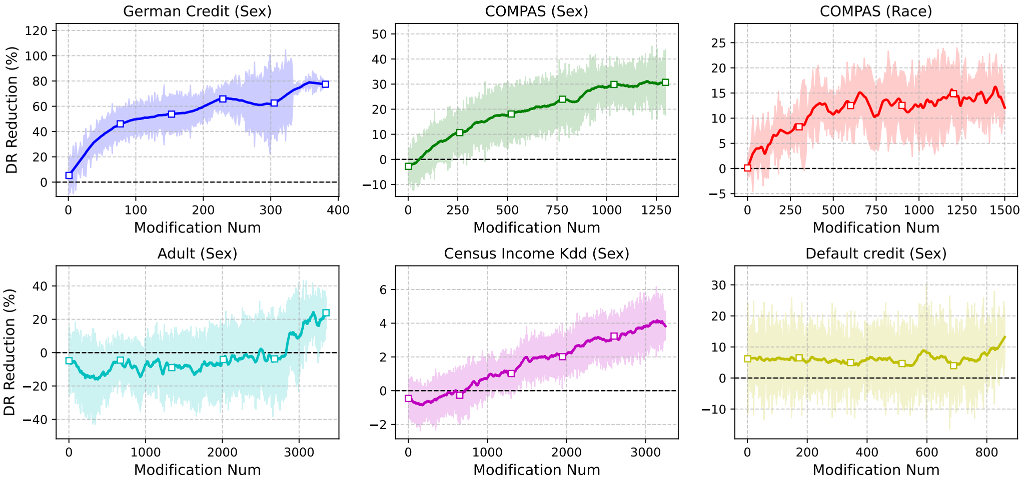
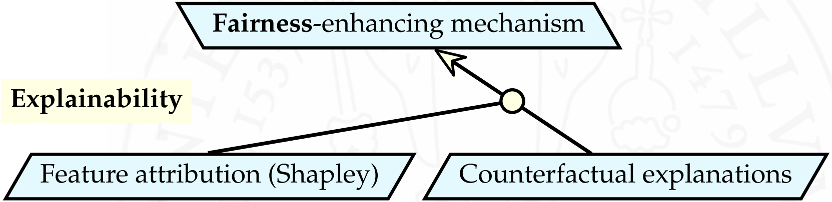


Figure 7. Percentage reduction in the discriminative risk (DR) across different datasets. The *x*-axis denotes the number of modifications applied (up to the maximum required under a fairness threshold $T = 0.05$), while the *y*-axis indicates the relative in DR, expressed as a percentage of the original value.

FairSHAP



FairSHAP: Preprocessing for Fairness Through Attribution-Based Data Augmentation[22]

- leverages Shapley value attribution to improve both individual and group fairness
- model-agnostic and transparent; is broadly applicable to tabular data, supports various models and SHAP algorithms, and can be seamlessly integrated into existing ML pipelines

[22] Zhu, Bian, and You, see n. 21.

Takeaway

- New fairness metrics are still needed to deal with multi-attribute, multi-valued, and realistic scenarios
- Prioritising individual fairness can offer stronger and more flexible leverage than focusing only on group fairness
- Connecting explainability with fairness makes mitigation more interpretable, targeted, and practical

Individual fairness

Lipschitz condition[23,24]

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 $(\breve{x}, a_1), (\breve{x}', a_1')$,

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

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. $\mathbb{P}\left(\frac{\boldsymbol{d}_y(h(\breve{x},a_1),h(\breve{x}',a_1'))}{\boldsymbol{d}_x((\breve{x},a_1),(\breve{x}',a_1'))} \geqslant \epsilon\right) \leqslant \delta$; or the $(\epsilon-\delta)$ language formulation: $\boldsymbol{d}_x((\breve{x}, a_1), (\breve{x}', a_1')) \leqslant \epsilon \Rightarrow \boldsymbol{d}_y(h(\breve{x}, a_1), h(\breve{x}', a_1')) \leqslant \delta$, 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.[25]

---

[23] Cynthia Dwork et al. "Fairness through awareness". In: *ITCS*. ITCS '12. Cambridge, Massachusetts: ACM, 2012, pp. 214–226. ISBN: 9781450311151.

[24] Additionally, a predictor satisfies individual fairness (Pratik Gajane and Mykola Pechenizkiy. "On formalizing fairness in prediction with machine learning". In: *FAT/ML*. 2018) iff: $h(\breve{x}, a_1) \approx h(\breve{x}', a_1') \mid \boldsymbol{d}_x((\breve{x}, a_1), (\breve{x}', a_1')) \approx 0$, 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.

[25] 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[26] and the **Theil index**[27]

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 $b_i = f(\breve{x}_i, a_{1i}) - y_i + 1$ and $\mu = \sum_i b_i / n$.

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.

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

[27] Haas, see n. 20.