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

Anonymous submission

Abstract

Increasing reliance on artificial intelligence and machine learning in high-stakes domains raises acute concerns about discrimination. However, existing fairness metrics remain constrained with limited applicability, especially when facing complex real-world scenarios beyond a single protected attribute. In this work, we take widely used group fairness measures as examples and comprehensively analyse their effectiveness in comparison with individual fairness. The empirical results illustrate that: (1) discrimination would be underestimated via binarisation; (2) traversal-based generalisation incurs computational cost; and (3) intersectional attributes remain superficially addressed.

1 Introduction

More and more concerns about the trustworthiness of artificial intelligence (AI) and machine learning (ML) systems are being raised due to the growing integration of them into socially consequential domains, including healthcare, recruitment, and jurisdiction (Liang et al. 2022; Chen et al. 2023; Hu et al. 2024; Glickman and Sharot 2024; Jones et al. 2024). Researchers working on algorithmic fairness (or fairness in machine learning, FairML) are devoted to mitigating the risks of ML algorithms or models reinforcing unjust human biases, with a broader goal of fostering more equitable decision-making. Despite substantial progress made in this field, however, existing fairness measures fall short in complex real-world scenarios because most of them remain limited to one single, typically binary, protected attribute (or sensitive attribute). This reductionism systematically underestimates disparities when sensitive attributes are multi-valued or intersectional. Attempts at incremental extensions—through binarisation or exhaustive traversal—either oversimplify or introduce prohibitive computational costs, resulting in fairness assessments that remain misaligned with the forms of discrimination encountered in real-world settings.

In this work, we comprehensively analyse and present the limited applicability of several commonly used fairness measures; we also challenge prevailing views on the incompatibility between accuracy and fairness as well as that among fairness measures themselves. We believe these empirical findings will offer interesting insights into the design of fairness measures that are more viable in practice.

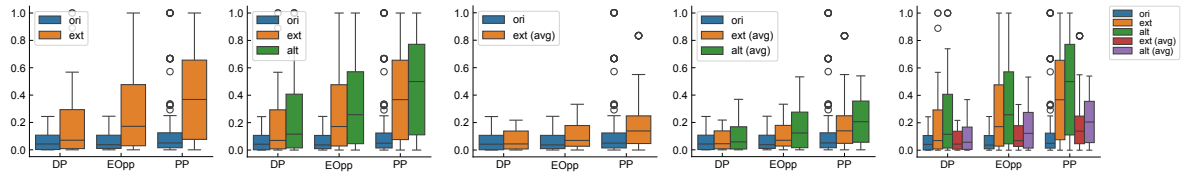
2 Preliminary

Three canonical criteria—independence, separation, and sufficiency—are widely recognised as cornerstones of formal definitions of statistical non-discrimination (Barocas, Hardt, and Narayanan 2023), respectively corresponding to requiring predictions to be independent of protected attributes ($\mathsf{A} \perp\!\!\!\perp \mathsf{R}$), conditionally independent given outcomes ($\mathsf{R} \perp\!\!\!\perp \mathsf{A} \mid \mathsf{Y}$), or that true labels be independent of protected attributes given predictions ($\mathsf{Y} \perp\!\!\!\perp \mathsf{A} \mid \mathsf{R}$). These conditions are rarely achievable simultaneously except in trivial settings, and even when satisfied, cannot guarantee the elimination of discrimination against individuals. For instance, equal hiring rates across groups may still disadvantage one group if selection quality differs due to structural imbalances in training data, like the so-called glass cliff. This illustrates that meeting a statistical condition is not equivalent to ensuring substantive fairness.

Before analysis, we list some typical fairness definitions and measures[1] in Appendix A, including four commonly used group fairness measures and individual fairness measures[2] for comparison. Note that we did not compare with the famous counterfactual fairness (Kusner et al. 2017; Gajane and Pechenizkiy 2018) because it is not a quantitative measure by definition, and more importantly, it heavily relies on causal inference, which means the analysis can only proceed with causal graphs as the learning model, limiting its applicability. We then provide the empirical analysis in the following, with experimental setups and additional results provided in Appendix B and C.

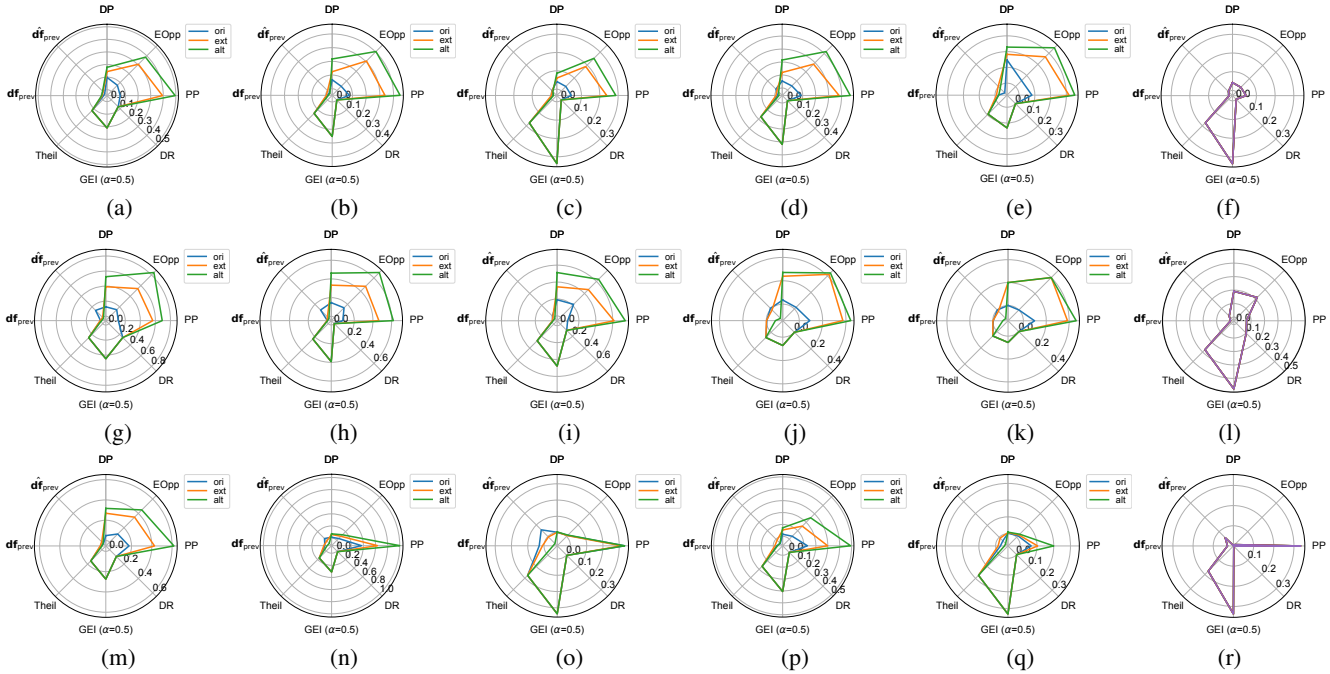
[1]Typical group fairness measures include demographic parity (DP) (Gajane and Pechenizkiy 2018; Jiang et al. 2020) (aka. statistical parity (Dwork et al. 2012; Chouldechova 2017)), equalised odds (EO) (Hardt, Price, and Srebro 2016; Haas 2019), equality of opportunity (EOpp) (Hardt, Price, and Srebro 2016; Gajane and Pechenizkiy 2018; Haas 2019), and predictive parity (PP) (Chouldechova 2017; Verma and Rubin 2018).

[2]The individual fairness measures are general entropy indices (GEI) (Speicher et al. 2018), the Theil index (Theil) (Haas 2019), as well as two other measures that can assess fairness from both individual and group aspects, that is, discriminative risk (DR) (Bian and Zhang 2023), and harmonic fairness via manifold (HFM) (Bian and Luo 2024; Bian, Luo, and Xu 2024).



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

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 their original definitions and the first two extension forms, analogously to Eq. (1), (3), and (4); note that (2) in binarisation is equivalent to (1); (c–d) Comparison between their original definitions and the last two extension forms, analogously to Eq. (2), (5), and (6); (e) Comparison between their original definitions and all four extension formulas.



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

(g) (h) (i) (j) (k) (l)

(m) (n) (o) (p) (q) (r)

Figure 2: Comparison of fairness measures between their original definitions and their corresponding extension formulas, on the Income (first row), Compas PPR (second row), and Compas PPVR (third row) datasets. Note that the first five columns use bagging, AdaBoost, LightGBM, AdaFair (trained using the first sensitive attribute), and AdaFair (trained using the second sensitive attribute), respectively; the sixth column uses LightGBM.

3 Empirical Analysis

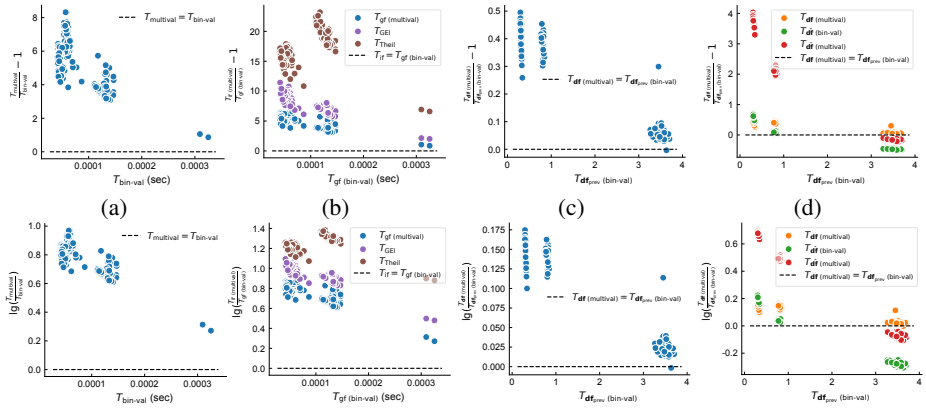
3.1 Illustrations on the insufficiency of existing fairness measures

As we can see from Table 1 in Appendix A, most quantitative fairness measures are designed with a single, typically binary, sensitive attribute in mind. While several extensions allow for applications to multi-valued attributes, our experiments reveal substantive risks in such adaptations, underscoring the need for fairness measures that are carefully constructed to address non-binary and intersectional attributes.

Binarisation underestimates discrimination As shown in Figure 1, binarising a multi-valued sensitive attribute (e.g. (2) and analogues) systematically yields smaller values compared with extensions such as (3) and (4). Even average-based forms in (5) and (6) often detect stronger disparities than binarisation. This pattern recurs across datasets regardless of the learning algorithms in use (shown in Figure 2), and suggests that binarisation oversimplifies the structure of disadvantage and can systematically underestimate discrimination. In contrast, individual fairness measures that naturally accommodate multi-valued attributes (such as GEI, Theil, and DR) remain stable across these settings. Given that even modest underestimations of bias can accumulate and amplify improper prejudice in human-AI interactions (Glickman and Sharot 2024), accurate measurement is essential for responsible deployment of ML systems.

Traversal-based generalisation incurs computational cost Extending group fairness measures (i.e. DP, EOpp, and PP) to multi-valued sensitive attributes is not only conceptually challenging but also computationally demanding.

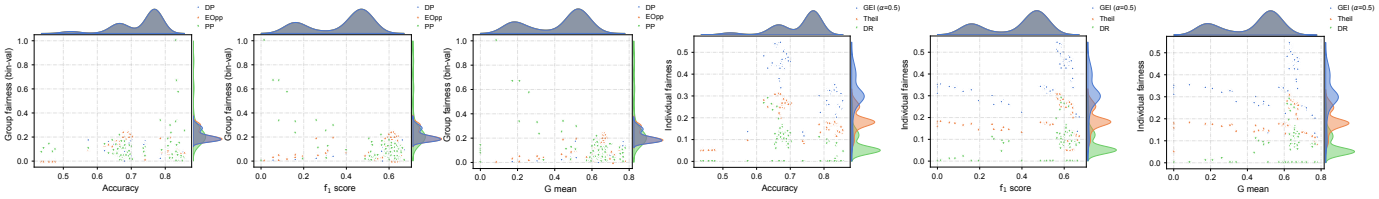
Figures (a) and (e) show that formulas such as Eq. (3) and (4) require almost an order of magnitude more time than their binarised counterparts like Eq. (2) in such multi-valued scenarios, and note that this is only for one 5- or 6-valued sensitive attribute. As for individual fairness measures that can handle one multi-valued sensitive attribute (e.g. general entropy indices (GEI) (Speicher et al. 2018) and the Theil



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

(e) (f) (g) (h)

Figure 3: Time cost comparison at different scales, for Income, Compas PPR, and Compas PPVR datasets. (a) and (e) Comparison of three commonly used group fairness measures and their extensions; (b) and (f) Comparison between group and individual fairness (GEI and Theil); (c–d) and (g–h) Comparison of HFM for binary-value and multi-value cases, where (c) and (g) are results via direct computation only, and (d) and (h) include results obtained by approximation algorithms (Bian and Luo 2024; Bian, Luo, and Xu 2024).



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

Figure 4: Scatter plot between performance (accuracy, f_1 score, or geometric mean (Akosa 2017), respectively) and fairness. Note that on the y-axis, the smaller the better; on the x-axis, the larger the better. (a–c) Using three commonly used group fairness measures, equivalent to (2) and its analogous formulas; (d–f) Using individual fairness measures.

index (Theil) (Haas 2019)), they have an even heavier computational burden, doubling to quadrupling that of the extensions of group fairness, presented in Figures (b) and (f). Similar observations that handling multi-valued cases brings about more computational costs recur in the remaining subfigures in Figure 3: Even for HFM, its maximum and average versions (Bian, Luo, and Xu 2024) increase runtime by up to 1.5× compared with its previous bi-valued version (Bian and Luo 2024), presented in Figures (c) and (g); Approximation strategies offer partial relief, presented in Figures (d) and (h), yet still lag behind simpler formulations.

Furthermore, it is obvious that degenerating intersectional attributes into one "super" discrete sensitive attribute through preprocessing is not an efficient way to handle them. For instance, consider data with two sensitive attributes $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 = \mathbb{Z}^{n_{a_1}} \times \mathbb{Z}^{n_{a_2}}$, where $n_{a_1}, n_{a_2} \geqslant 2$. After preprocessing, the combined single sensitive attribute would be $\mathcal{A}' = \mathbb{Z}^{n_{a_1} \times n_{a_2}}$. This approach 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 $\mathbb{Z}^{12}$). Similarly, if the number of sensitive attributes itself increases, the computational burden also rises (e.g. $\mathcal{A}'$ could shift from $\mathbb{Z}^{2\times 6} = \mathbb{Z}^{12}$ to $\mathbb{Z}^{2\times 6\times 3} = \mathbb{Z}^{36}$). The larger the number of values for each sensitive attribute, the longer it takes to compute. In such cases, the computational cost may be reduced if one fairness measure can handle multiple sensitive attributes directly. This approach can be viewed as decomposing the original complex problem into smaller and more manageable sub-problems, akin to the divide-and-conquer strategy.

These results indicate that efficient fairness assessment for multi-valued and even intersectional attributes remains an open practical challenge, motivating the development of more computationally tractable approaches.

3.2 Illustrations on incompatibility views

Accuracy and fairness are not strictly incompatible A prevailing view in the literature describes fairness and accuracy as fundamentally at odds, with improvements in one often assumed to degrade the other (Berk et al. 2021). Our experiments shown in Figures 4 and 6 partly support this view: when models operate at low to moderate accuracy, gains in predictive performance can indeed exacerbate disparities. However, the relationship is more nuanced. At sufficiently high levels of accuracy, improvements can coincide with enhanced fairness, particularly in the case of individual fairness, demonstrated in Figures (d) to (f). For group fairness, this alignment is less obvious in the original formulations, but becomes clearer in their extended forms presented in Figures (a) to (j), as well as the remaining subfigures of Figure 6 in Appendix C. These findings suggest that the long-assumed trade-off is not universal; instead, it points to opportunities for promoting fairness and accuracy

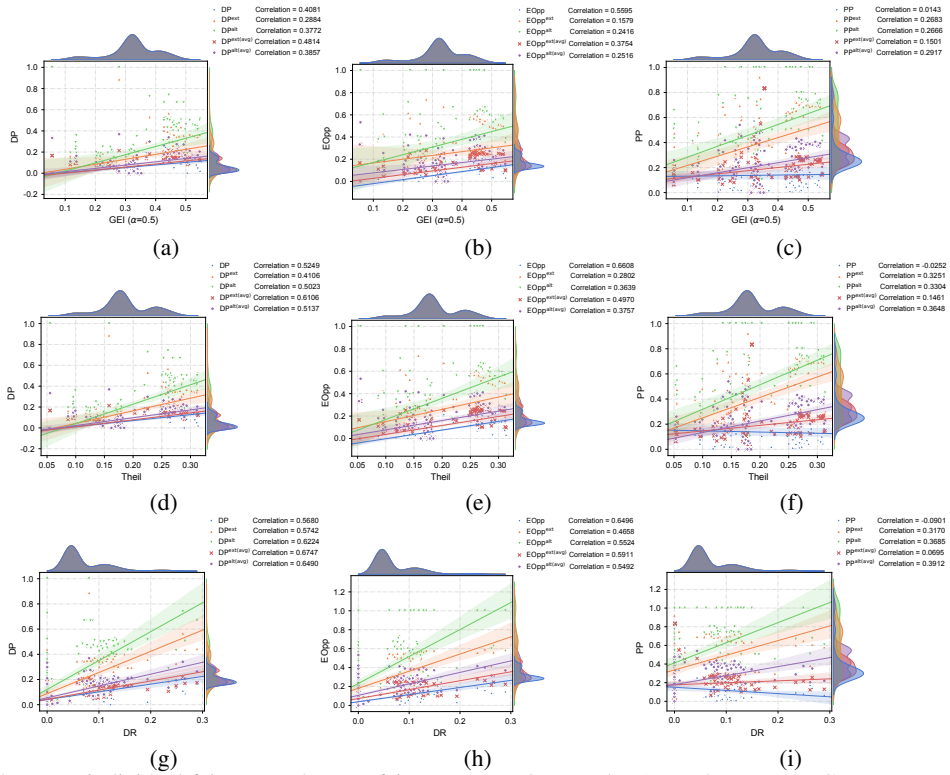


Figure 5: Relation between individual fairness and group fairness (DP, EOpp, and PP), on the Income, Compas PPR, and Compas PPVR datasets. Note that on both x- and y-axes, the smaller the better. (a–c), (d–f), and (g–i) use GEI, the Theil index, and DR, respectively, as the individual fairness measure.

jointly, especially by prioritising individual fairness in high-performance regimes.

Individual fairness and group fairness are not inherently incompatible In the literature, group fairness is frequently framed as incompatible with individual fairness (Hardt, Price, and Srebro 2016; Pleiss et al. 2017): for example, satisfying group fairness still permits predicting qualified applicants in one demographic but random individuals in another. Yet the reverse may not be the case: Even intuitively, if individuals are all treated fairly, the fair treatment concerning their demographic is supposed to follow, meaning that the satisfaction of individual fairness should achieve group fairness as well. This intuition is empirically demonstrated in Figures 5(d), 5(e), 5(g), and 5(h): The worsen individual fairness (Theil and DR) signifies the same worsen tendency of group fairness (DP and EOpp), as well as analogously when individual fairness gets better. The same trend of change is also evidenced by their correlations: Figures 5(d) and 5(e) show that Theil is moderately correlated with DP, EOpp, and the average forms of DP's extensions; Figures 5(g) and 5(h) show that DR is moderately and nearly highly correlated with DP, EOpp, and the average forms of their corresponding extensions.

Besides, neither Theil nor DR shows a high correlation with PP in Figures 5(f) and 5(i), which also supports the incompatibility among the three statistical non-discrimination criteria; similar observation can also be found in the average HFM by Figures 7(g) to 7(i). The incompatibility among these criteria is also supported by Figures 7(a) to 7(f), where the previous and maximum HFM have moderate to nearly high correlation with PP, yet are almost irrelevant to DP and EOpp. Note that HFM, differing from all other fairness measures, can only capture the extra bias introduced in the learning procedure; the reason why the average HFM exhibits different correlations with group fairness may be that the average HFM among the three versions is not as extreme as its two counterparts using the maximal operator, therefore, more consistent with discrimination evaluated via predictions. As for GEI, it shows a moderate correlation with EOpp only, presented in Figure 5(b), and therefore has fewer values as a signal compared to Theil and DR.

4 Conclusion

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

Agarwal, A.; Dudík, M.; and Wu, Z. S. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *ICML*, volume 97, 120–129. PMLR.

Akosa, J. 2017. Predictive accuracy: A misleading performance measure for highly imbalanced data. In *Proceedings*

of the SAS global forum, volume 12, 1–4. SAS Institute Inc. Cary, NC, USA.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. Cambridge, MA, USA: MIT Press.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociol Methods Res*, 50(1): 3–44.

Bian, Y.; and Luo, Y. 2024. Does machine bring in extra bias in learning? Approximating fairness in models promptly. *arXiv preprint arXiv:2405.09251*.

Bian, Y.; Luo, Y.; and Xu, P. 2024. Approximating discrimination within models when faced with several non-binary sensitive attributes. *arXiv preprint arXiv:2408.06099*. Under review.

Bian, Y.; and Zhang, K. 2023. Increasing fairness via combination with learning guarantees. *arXiv preprint arXiv:2301.10813*. Under review.

Boeschoten, L.; van Kesteren, E.-J.; Bagheri, A.; and Oberski, D. L. 2021. Achieving fair inference using error-prone outcomes. *Int J Interact Multimed Artif Intell*, 6(5).

Chen, R. J.; Wang, J. J.; Williamson, D. F.; Chen, T. Y.; Lipkova, J.; Lu, M. Y.; Sahai, S.; and Mahmood, F. 2023. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng*, 7(6): 719–742.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2): 153–163.

Compas. ???? Propublica-recidivism and Propublica-violent-recidivism datasets. [EB/OL]. Latest accessed August 29, 2022.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *SIGKDD*, 797–806. ACM. ISBN 9781450348874.

Credit. ???? Statlog (German credit data).

Cruz, A. F.; Belém, C.; Bravo, J.; Saleiro, P.; and Bizarro, P. 2023. FairGBM: Gradient Boosting with Fairness Constraints. In *ICLR*.

Diana, E.; Sharifi-Malvajerdi, S.; and Vakilian, A. 2024. Minimax group fairness in strategic classification. In *SatML*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *ITCS*, ITCS '12, 214–226. New York, NY, USA: ACM. ISBN 9781450311151.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *SIGKDD*, 259–268.

Foulds, J. R.; Islam, R.; Keya, K. N.; and Pan, S. 2020. An intersectional definition of fairness. In *ICDE*, 1918–1921. IEEE.

Gajane, P.; and Pechenizkiy, M. 2018. On formalizing fairness in prediction with machine learning. In *FAT/ML*.

Ghosh, A.; Genuit, L.; and Reagan, M. 2021. Characterizing intersectional group fairness with worst-case comparisons. In Lamba, D.; and Hsu, W. H., eds., *Proceedings of 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI)*, volume 142, 22–34. Virtual: PMLR.

Glickman, M.; and Sharot, T. 2024. How human-AI feedback loops alter human perceptual, emotional and social judgements. *Nat Hum Behav*.

Grgić-Hlača, N.; Zafar, M. B.; Gummadi, K. P.; and Weller, A. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, 11. Barcelona, Spain.

Grgić-Hlača, N.; Zafar, M. B.; Gummadi, K. P.; and Weller, A. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *AAAI*, volume 32.

Haas, C. 2019. The price of fairness - A framework to explore trade-offs in algorithmic fairness. In *ICIS*. Association for Information Systems.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *NIPS*, volume 29, 3323–3331. Red Hook, NY, USA: Curran Associates Inc.

Hu, T.; Kyrychenko, Y.; Rathje, S.; Collier, N.; van der Linden, S.; and Roozenbeek, J. 2024. Generative language models exhibit social identity biases. *Nat Comput Sci*, 1–11.

Income. ???? Adult.

Iosifidis, V.; and Ntoutsi, E. 2019. AdaFair: Cumulative fairness adaptive boosting. In *CIKM*, 781–790. New York, NY, USA: ACM.

Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; and Chiappa, S. 2020. Wasserstein fair classification. In *UAI*, 862–872. PMLR.

Jones, C.; Castro, D. C.; De Sousa Ribeiro, F.; Oktay, O.; McCradden, M.; and Glocker, B. 2024. A causal perspective on dataset bias in machine learning for medical imaging. *Nat Mach Intell*, 6(2): 138–146.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *NIPS*, volume 30, 3146–3154.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *ICML*, 2564–2572. PMLR.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2019. An empirical study of rich subgroup fairness for machine learning. In *FAT*, 100–109.

Kilbertus, N.; Rojas Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *NIPS*, volume 30.

Kim, M. P.; Ghorbani, A.; and Zou, J. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *AIES*, AIES '19, 247–254. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *NIPS*, volume 30, 4069–4079. NIPS Proceedings.

Liang, W.; Tadesse, G. A.; Ho, D.; Fei-Fei, L.; Zaharia, M.; Zhang, C.; and Zou, J. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nat Mach Intell*, 4(8): 669–677.

Luong, B. T.; Ruggieri, S.; and Turini, F. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *SIGKDD*, 502–510.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. In *NIPS*, volume 30.

Ricci. ???? Ricci: Firefighter promotion exam scores.

Speicher, T.; Heidari, H.; Grgic-Hlaca, N.; Gummadi, K. P.; Singla, A.; Weller, A.; and Zafar, M. B. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *SIGKDD*, 2239–2248.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *FairWare*, 1–7.

Wang, Z.; Huang, C.; and Yao, X. 2024. Procedural fairness in machine learning. *arXiv preprint arXiv:2404.01877*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, 1171–1180.

A Preliminary (cont.)

In the following, we list several commonly used group fairness measures, originally defined for one bi-valued sensitive attribute, and extend them to apply to non-binary cases. We also include some individual fairness measures for comparison. Note that a task relevant to discrimination or bias mitigation is usually binary classification or prediction using a score function in some cases, and for instance $(\mathbf{x}, a_1)$ where a_1 denotes one bi-valued sensitive attribute (i.e. $a_1 \in \mathcal{A}_1 = \{0, 1\}$), $a_1 = 1$ represents the privileged group, while $a_1 = 0$ represents the marginalised group.

Independence-based measures of group fairness Demographic parity (Gajane and Pechenizkiy 2018; Jiang et al. 2020) (DP, aka. statistical parity (Dwork et al. 2012; Chouldechova 2017)), that is,

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

along with related notions (such as disparate impact, disparate treatment, conditional statistical parity, and bounded group loss), instantiates the independence criterion. While mathematically straightforward, DP is limited in scope because its canonical form applies only to one single binary sensitive attribute. Extensions to multi-valued attributes are often achieved through binarisation, grouping all non-privileged categories together. That is to say, it can be evaluated by

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

where $a_1 \in \mathcal{A}_1 = \{1, 2, ..., n_{a_1}\}$, $a_1 = 1$ still denotes the privileged group, and $a_1 \neq 1$ denotes the marginalised groups. Note that (2) is equivalent to (1) when $\mathcal{A}_1 = \{0, 1\}$.

However, binarisation oversimplifies the discrimination complexity in reality and risks masking important disparities: for example, treating several marginalised groups as homogeneous may obscure intersectional harms. One refined extension, inspired by statistical parity (Corbett-Davies et al. 2017; Agarwal, Dudík, and Wu 2019) (aka. disparate treatment (Zafar et al. 2017; Corbett-Davies et al. 2017; Haas 2019))[3], can be defined as

$$\begin{aligned}\mathrm{DP}^{\text{ext}}(f) = \max_{j \in \mathcal{A}_1}\{|\ \ &\mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = j) \\ &-\mathbb{P}(f(\mathbf{x}, a_1) = 1)|\}\,,\end{aligned} \tag{3}$$

and another meticulous formula inspired by DP (Jiang et al. 2020) is

$$\begin{aligned}\mathrm{DP}^{\text{alt}}(f) = \max_{j,k \in \mathcal{A}_1, k \neq j}\{|\ \ &\mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = j) \\ &-\mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = k)|\}\,,\end{aligned} \tag{4}$$

[3]Disparate treatment (Zafar et al. 2017), also indicated as "statistical parity" in (Corbett-Davies et al. 2017; Haas 2019), is defined as $\mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = j) = \mathbb{P}(f(\mathbf{x}, a_1) = 1)\,,\ \forall j \in \{0, 1\}$. It is possible to apply to one multi-valued sensitive attribute (Corbett-Davies et al. 2017; Agarwal, Dudík, and Wu 2019), called statistical parity, that is, $\mathbb{P}(f(\mathbf{x}, a_1) = 1 \mid a_1 = j) = \mathbb{P}(f(\mathbf{x}, a_1) = 1)\,,\ \forall j \in \mathcal{A}_1 = \{1, 2, ..., n_{a_1}\}$.

as well as their corresponding average forms

$$\mathrm{DP}^{\text{ext(avg)}}(f) = \tfrac{1}{n_{a_1}} \textstyle\sum_{j\in\mathcal{A}_1}\{|\ \ \mathbb{P}(f(\mathbf{x},a_1)=1 \mid a_1=j) - \mathbb{P}(f(\mathbf{x},a_1)=1)|\}\,, \tag{5}$$

$$\mathrm{DP}^{\text{alt(avg)}}(f) = \tfrac{2}{n_{a_1}(n_{a_1}-1)} \textstyle\sum_{j=1}^{n_{a_1}-1}\sum_{k=j+1}^{n_{a_1}}\{| \mathbb{P}(f(\mathbf{x},a_1)=1|a_1=j) - \mathbb{P}(f(\mathbf{x},a_1)=1|a_1=k)|\}\,. \tag{6}$$

Although these attempts capture finer distinctions, they essentially still rely on treating intersectional attributes as a single discrete variable. As a result, intersectionality remains only superficially addressed.

Separation-based measures of group fairness Equality of opportunity (EOpp) (Hardt, Price, and Srebro 2016; Gajane and Pechenizkiy 2018; Haas 2019) defined as

$$\mathrm{EOpp}(f) = |\ \ \mathbb{P}(f(\mathbf{x},a_1)=1 \mid a_1=0, y=1) - \mathbb{P}(f(\mathbf{x},a_1)=1 \mid a_1=1, y=1)|\,, \tag{7}$$

as well as relevant measures (such as equalised odds (EO) (Hardt, Price, and Srebro 2016; Haas 2019) defined as

$$\begin{aligned}\mathrm{EO} = \tfrac{1}{2}[|\ \ &\mathbb{P}(f(\mathbf{x},a_1)=1 \mid a_1=0, y=0) - \mathbb{P}(f(\mathbf{x},a_1)=1 \mid a_1=1, y=0)| \\ +|\ \ &\mathbb{P}(f(\mathbf{x},a_1)=1 \mid a_1=0, y=1) - \mathbb{P}(f(\mathbf{x},a_1)=1 \mid a_1=1, y=1)|]\,,\end{aligned}$$

predictive equality, and γ-subgroup fairness), instantiates the separation criterion. They require that all groups experience the same false negative rate, which is associated with denied opportunities when acceptance is desired. Like DP, EOpp generalises imperfectly to multi-valued or intersectional settings, such as

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

$$\mathrm{EOpp}^{\text{ext}}(f) = \max_{j\in\mathcal{A}_1}\{|\mathbb{P}(f(\mathbf{x},a_1)=1 \mid a_1=j, y=1) - \mathbb{P}(f(\mathbf{x},a_1)=1 \mid y=1)|\}\,,$$

$$\mathrm{EOpp}^{\text{alt}}(f) = \max_{j,k\in\mathcal{A}_1, k\neq j}\{|\mathbb{P}(f(\mathbf{x},a_1)=1 \mid a_1=j, y=1) - \mathbb{P}(f(\mathbf{x},a_1)=1 \mid a_1=k, y=1)|\}\,,$$

as well as $\mathrm{EOpp}^{\text{ext(avg)}}$ and $\mathrm{EOpp}^{\text{alt(avg)}}$ analogous with Eq. (5) to (6), and binarisation and its extended formulas remain constrained.

Sufficiency-based measure of group fairness Predictive parity (PP) (Chouldechova 2017; Verma and Rubin 2018), defined as

$$\mathrm{PP}(f) = |\ \ \mathbb{P}(y=1 \mid a_1=0, f(\mathbf{x},a_1)=1) - \mathbb{P}(y=1 \mid a_1=1, f(\mathbf{x},a_1)=1)|\,, \tag{9}$$

is closely aligned with calibration and satisfies the sufficiency criterion. Extensions like

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

$$\mathrm{PP}^{\text{ext}}(f) = \max_{j\in\mathcal{A}_1}\{|\mathbb{P}(y=1 \mid a_1=j, f(\mathbf{x},a_1)=1) - \mathbb{P}(y=1 \mid f(\mathbf{x},a_1)=1)|\}\,,$$

$$\mathrm{PP}^{\text{alt}}(f) = \max_{j,k\in\mathcal{A}_1, k\neq j}\{|\mathbb{P}(y=1 \mid a_1=j, f(\mathbf{x},a_1) = 1) - \mathbb{P}(y=1 \mid a_1=k, f(\mathbf{x},a_1)=1)|\}\,,$$

as well as $\mathrm{PP}^{\text{ext(avg)}}$ and $\mathrm{PP}^{\text{alt(avg)}}$ analogous with Eq. (5) to (6), again mirror those of DP and EOpp, allowing application to multi-valued attributes but inheriting the same limitations concerning intersectionality.

Individual fairness in comparison The Lipschitz condition (Dwork et al. 2012) is primarily viewed as individual fairness yet not a quantitative measure. It means that a mapping or 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 $(\mathbf{x},a_1),(\mathbf{x}',a_1')$,

$$\boldsymbol{d}_y(h(\mathbf{x},a_1),h(\mathbf{x}',a_1')) \leqslant \lambda\cdot\boldsymbol{d}_x((\mathbf{x},a_1),(\mathbf{x}',a_1'))\,, \tag{11}$$

where $\boldsymbol{d}_y$ and $\boldsymbol{d}_x$ are (task-specific) distance metrics. It can also be written as the probability Lipschitzness, i.e.

$$\mathbb{P}\left(\tfrac{\boldsymbol{d}_y(h(\mathbf{x},a_1),h(\mathbf{x}',a_1'))}{\boldsymbol{d}_x((\mathbf{x},a_1),(\mathbf{x}',a_1'))} \geqslant \epsilon\right) \leqslant \delta\,;$$

or the $(\epsilon-\delta)$ language formulation: $\boldsymbol{d}_x((\mathbf{x},a_1),(\mathbf{x}',a_1')) \leqslant \epsilon \Rightarrow \boldsymbol{d}_y(h(\mathbf{x},a_1),h(\mathbf{x}',a_1')) \leqslant \delta$, where $\epsilon\geqslant 0$ and $\delta\geqslant 0$. Note that λ is a positive constant. Additionally, in Gajane and Pechenizkiy (2018), a predictor satisfies individual fairness if and only if: $h(\mathbf{x},a_1) \approx h(\mathbf{x}',a_1') \mid \boldsymbol{d}_x((\mathbf{x},a_1),(\mathbf{x}',a_1')) \approx 0$, 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. 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 (Luong, Ruggieri, and Turini 2011; Boeschoten et al. 2021).

Except for the Lipschitz condition, we list two individual fairness measures (i.e. general entropy indices (Speicher et al. 2018) and the Theil index (Haas 2019) and two other fairness measures that can assess fairness from both individual and group aspects.

Definition 1 (General entropy indices and the Theil index)**.** *For a constant $\alpha\notin\{0,1\}$, the generalised entropy indices for a problem with n instances are defined, to quantify algorithmic unfairness, as*

$$\mathrm{GEI}^{\alpha} = \tfrac{1}{n\alpha(\alpha-1)}\textstyle\sum_{i=1}^{n}\left(\left(\frac{b_i}{\mu}\right)^{\alpha}-1\right), \tag{12}$$

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

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

$$\mathrm{Theil} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{b_i}{\mu}\log\left(\frac{b_i}{\mu}\right). \tag{13}$$

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

Definition 2 (Discriminative risk (Bian and Zhang 2023))**.**

$$\mathrm{DR}(f) = \mathbb{E}[\mathbb{I}(f(\mathbf{x},\mathbf{a})\neq f(\mathbf{x},\tilde{\boldsymbol{a}}))]\,, \tag{14}$$

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

Definition 3 (Harmonic fairness via manifold)**.** *Given a dataset $D = (\mathsf{X},\mathsf{A},\mathsf{Y})$, it (abbreviated to HFM) (Bian and Luo 2024; Bian, Luo, and Xu 2024) has three versions: (1) the* previous *HFM for one bi-valued sensitive attribute, and (2) the* maximal (resp. average) *HFM for several multi-valued sensitive attributes.*

For one bi-valued SA $a_1\in\mathcal{A}_1=\{0,1\}$, D is divided into $D_1 = \{(\mathbf{x}_a,y)\triangleq(\mathbf{x},a_1,y)\in D \mid a_1=1\}$ 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

$$\mathrm{HFM}_{prev}(f) = \boldsymbol{g}_f(D_1,\bar{D}_1)/\boldsymbol{g}(D_1,\bar{D}_1) - 1\,, \tag{15}$$

where

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

Table 1: Summary of existing fairness measures. Note that '#sen-att' denotes the number of sensitive attributes (i.e. n_a), and mark $*$ indicates it is one of the *distributive* fairness measures.

Name of *measure*	Fairness type	Meaning		Applicable situation(s) in definition			Non-binary handling[3]	
		quant.[1]	fairer	#label (n_c)	#sen-att (n_a)[2]	#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 (Feldman et al. 2015; Zafar et al. 2017)	*, group-	yes	*larger* value	binary	singular	bi-valued	yes	indirectly
Disparate treatment (Zafar et al. 2017)	*, group-	poss	lower value	binary	singular	bi- (multi- allowed)	yes	indirectly
Conditional statistical parity (Corbett-Davies et al. 2017)	*, group-	poss	lower value	binary	singular	multi-valued	—	indirectly
Bounded group loss (Agarwal, Dudík, and Wu 2019)	*, group-	poss	lower value	binary	singular	multi-valued	—	indirectly
Strategic minimax fairness (Diana, Sharifi-Malvajerdi, and Vakilian 2024)	*, group-	no	—	bi-/multi-	singular	multi-valued	—	indirectly
Equalised odds (Hardt, Price, and Srebro 2016; Haas 2019)	*, 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 (Corbett-Davies et al. 2017)	*, group-	poss	lower value	binary	singular	multi-valued	—	indirectly
γ-subgroup fairness (Kearns et al. 2018, 2019)	*, group-	yes	lower value	binary	singular	bi-valued	yes	indirectly
Predictive parity (Chouldechova 2017; Verma and Rubin 2018)	*, group-	yes	lower value	binary	singular	bi-valued	yes	indirectly
Lipschitz condition (Dwork et al. 2012; Gajane and Pechenizkiy 2018)	*, individual-	no	—	binary	singular	bi-valued	yes	indirectly
General entropy indices (Speicher et al. 2018) (and the Theil index (Haas 2019))	*, individual-	yes	lower value	binary	singular	multi-valued	—	indirectly
Counterfactual fairness (Kusner et al. 2017; Gajane and Pechenizkiy 2018)	*, individual-	no	—	binary	allows plural	multi- allowed	yes	indirectly
Proxy discrimination (Kilbertus et al. 2017)	*, individual-	no	—	binary	singular	multi- allowed	yes	indirectly
Discriminative risk (Bian and Zhang 2023)	*, [4]	yes	lower value	bi-/multi-	allows plural	multi-valued	—	—
Harmonic fairness via manifold	*, [4]	yes	lower value	bi-/multi-	allows plural	multi-valued	—	—
Multiaccuracy (Kim, Ghorbani, and Zou 2019)	*,group-	poss	lower value	binary	singular	multi-valued	—	indirectly
Differentially fair (Foulds et al. 2020)	*,group-	poss	—	binary	allows plural	bi- (multi- allowed)	yes	indirectly
Group benefit ratio and worst-case min-max ratio (Ghosh, Genuit, and Reagan 2021)	*,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 (Wang, Huang, and Yao 2024)	procedural	yes	lower value	binary	singular	bi-valued	yes	indirectly

[1] Whether the corresponding fairness is a *quantitative* measure. Note that many (including DP, EO, EOpp, and PP) may not be quantitative by definition, but remain a possibility to be one, indicated by 'poss'. DP, EO, EOpp, and PP are indicated as 'yes' as there are widely used forms for them.
[2] Whether it applies to intersectional SAs, in other words, 'singular' and 'plural' mean it can handle one sensitive attribute only and more than one sensitive attribute, respectively.
[3] Whether it can handle non-binary cases or whether it is possible to be extended. Note that 'indirectly' means essentially it requires intersectional attributes to be handled as one discrete- (or multi-)valued sensitive attribute.
[4] Discriminative risk (DR) is viewed as individual fairness primarily but can reflect group-level fairness as well; Harmonic fairness via manifold (HFM) (Bian and Luo 2024; Bian, Luo, and Xu 2024) is based on distances between individuals and that between sets, and thus can reflect both individual- and group-level fairness.
[5] All three initial procedural fairness measures (Grgić-Hlača et al. 2016, 2018) depend critically on member judgements regarding whether the use of particular features in decision-making is discriminatory. These member judgements may change over time, resulting in unstable outcomes and unpredictable computational demands as systems are recalibrated repeatedly.

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

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

$$\mathrm{HFM}(f) = \log\left(\boldsymbol{g}_{f,\mathbf{a}}(D)/\boldsymbol{g}_{\mathbf{a}}(D)\right), \tag{16a}$$

$$\mathrm{HFM}^{avg}(f) = \log\left(\boldsymbol{g}^{avg}_{f,\mathbf{a}}(D)/\boldsymbol{g}^{avg}_{\mathbf{a}}(D)\right), \tag{16b}$$

where

$$\boldsymbol{g}_{\cdot,\mathbf{a}}(D; \ddot{y}) = \max_{1\leqslant i\leqslant n_a} \boldsymbol{g}_{\cdot,\mathbf{a}}(D, a_i; \ddot{y}), \tag{17a}$$

$$\boldsymbol{g}^{avg}_{\cdot,\mathbf{a}}(D; \ddot{y}) = \tfrac{1}{n_a}\textstyle\sum_{i=1}^{n_a} \boldsymbol{g}^{avg}_{\cdot,\mathbf{a}}(D, a_i; \ddot{y}), \tag{17b}$$

and

$$\boldsymbol{g}_{\cdot,\mathbf{a}}(D, a_i; \ddot{y}) = \max_{j\in\{1,2,\ldots,n_{a_i}\}}\{ \max_{(\mathbf{x}_a,y)\in D_j} \min_{(\mathbf{x}'_a,y')\in \bar{D}_j} \boldsymbol{d}((\mathbf{x}, \ddot{y}), (\mathbf{x}', \ddot{y}'))\},$$

$$\boldsymbol{g}^{avg}_{\cdot,\mathbf{a}}(D, a_i; \ddot{y}) = \tfrac{1}{n}\textstyle\sum_{j\in\{1,2,\ldots,n_{a_i}\}} \sum_{(\mathbf{x},y)\in D_j} \min_{(\mathbf{x}',y')\in \bar{D}_j} \boldsymbol{d}((\mathbf{x}, \ddot{y}), (\mathbf{x}', \ddot{y}')).$$

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

B Experimental Setup

In this section, we elaborate on our experimental settings to evaluate existing fairness measures and their possible extensions.

Datasets We collected five public datasets but mainly used three of them in the experiments, because Ricci and Credit only have bi-valued sensitive attributes. Detailed information about them are provided in Table 2.

Table 2: Data statistics.

Dataset	#inst[1]	#feat[1]		1st sen-att[2]			2nd sen-att[2]		
		raw	prep		#val	#in-priv		#val	#in-priv
Ricci (Ricci)	118	5	6	race	3	68	—	—	—
Credit (Credit)	1000	20	58	sex	2	690	age	2	851
Income (Income)	30162	13	98	race	5	25933	sex	2	20380
Compas PPR [3]	6167	10	401	sex	2	4994	race	6	2100
Compas PPVR[3]	4010	10	327	sex	2	3173	race	6	1452

[1] The columns '#inst' and '#feat' represent the number of instances and the number of features (including one or two sensitive attributes, but excluding classification labels), respectively. Note that 'prep' is the number of features after preprocessing.
[2] For each sensitive attribute (sen-att), '#val' and '#in-priv' mean the number of its values and the number of members in the privileged group, respectively.
[3] Compas PPR and PPVR datasets come from (Compas).

Evaluation metrics We consider accuracy and f_1 score as performance metrics. We also consider the geometric mean (Akosa 2017) because data imbalance usually occurs within the datasets relevant to discrimination/bias mitigation. We directly use the time cost as the efficiency metric. As for fairness measures, we choose three commonly used group fairness measures, some individual fairness measures, as well as harmonic fairness via manifold (Bian and Luo 2024; Bian, Luo, and Xu 2024): (1) the three commonly used group fairness measures are demographic parity (DP) (Feldman et al. 2015; Gajane and Pechenizkiy 2018), equality of opportunity (EOpp) (Hardt, Price, and Srebro 2016), and predictive parity (PP) (Chouldechova 2017; Verma and Rubin 2018); (2) the individual fairness measures used are general entropy indices (GEI) (Speicher et al. 2018) and the Theil index (Haas 2019); and (3) two other measures that can assess fairness from both individual and group aspects, that is, discriminative risk (DR) (Bian and Zhang 2023) and harmonic

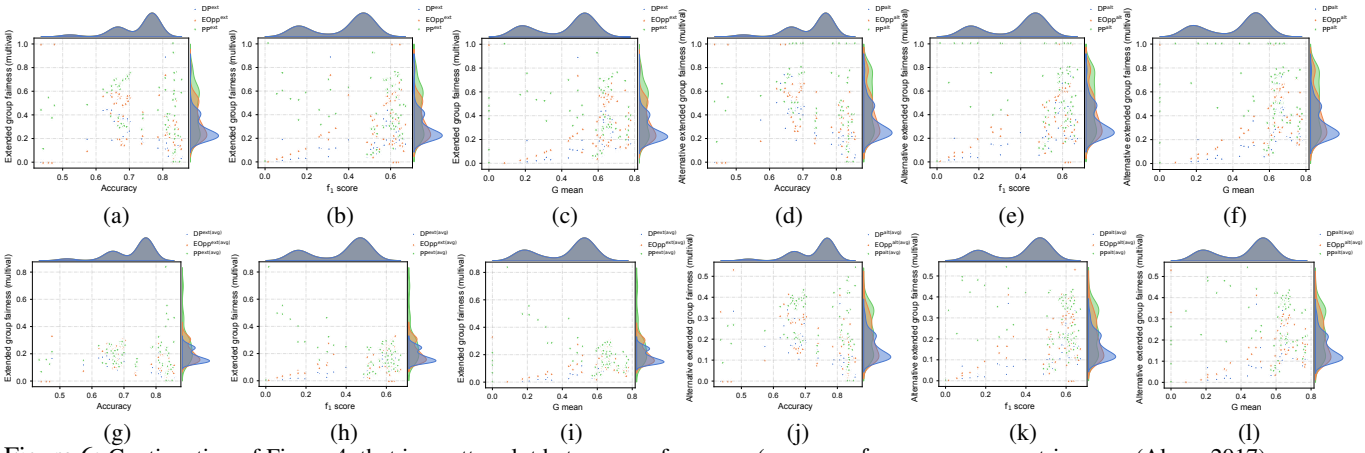


Figure 6: Continuation of Figure 4, that is, scatter plot between performance (accuracy, f_1 score, or geometric mean (Akosa 2017), respectively) and fairness. (a–c) Using Eq. (3) and its analogous forms; (d–f) Using Eq. (4) and its analogous forms; (g–i) Using Eq. (5) and its analogous forms; (j–l) Using Eq. (6) and its analogous forms.

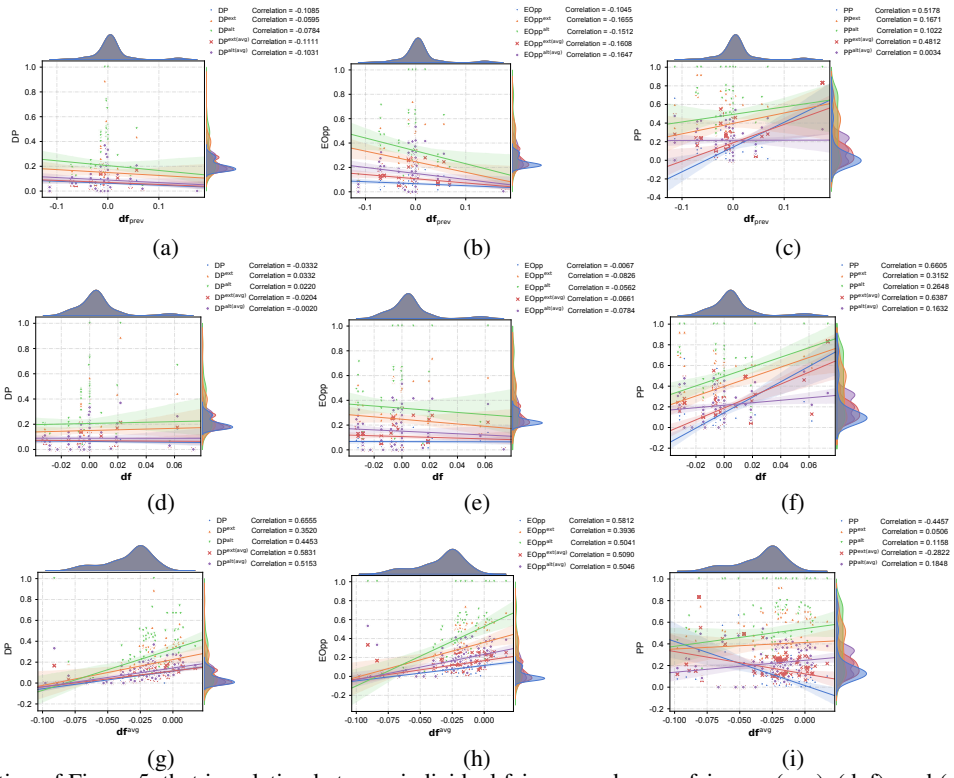


Figure 7: Continuation of Figure 5, that is, relation between individual fairness and group fairness. (a–c), (d–f), and (g–i) use the previous HFM (Bian and Luo 2024), the maximum HFM (Bian, Luo, and Xu 2024), and the average HFM (Bian, Luo, and Xu 2024), respectively, as the individual fairness measure.

fairness via manifold (HFM, with three versions) (Bian and Luo 2024; Bian, Luo, and Xu 2024).

Implementation details We mainly use bagging, AdaBoost, LightGBM (Ke et al. 2017), FairGBM (Cruz et al. 2023), and AdaFair (Iosifidis and Ntoutsi 2019) as learning algorithms, where FairGBM and AdaFair are two fairness-aware ensemble-based methods. Standard 5-fold cross-validation is used in these experiments; in other words, in each iteration, the entire dataset is divided into two parts, with 80% as the training set and 20% as the test set. Also, features of datasets are scaled in preprocessing to lie between 0 and 1.

C Additional Empirical Results

We include more empirical results here to save space, that is, Figures 6 to 7.