Beyond Base Metric: The Critical and Overlooked Role of Meta-Metrics in Intersectional Fairness

JEANNE MONNIER, EURECOM, France and Orange Research, France
THOMAS GEORGE, Orange Research, France
FRÉDÉRIC GUYARD, Orange Research, France
CHRISTÈLE TARNEC, Orange Research, France
MARIOS KOUNTOURIS, EURECOM, France and University of Granada, Spain

Fairness assessment is essential to align AI systems with normative and societal values. Because fairness is an inherently contested ethical concept, it is difficult to measure. Although most fairness metrics were developed for binary classification settings, real-world deployments require attention to intersectional social categories [10]. In these settings, an additional and often overlooked modeling step arises: aggregating subgroup-level measures and multiple pairwise group disparities into a single summary score. This step requires selecting a meta-metric, and different choices can produce divergent or even conflicting fairness assessments. We characterize the space of such choices through a three-step taxonomy that makes explicit the normative and technical assumptions embedded in meta-metric design. We show that different design decisions produce substantially different fairness evaluations, with important ethical, methodological, and practical consequences. We therefore argue that aggregation choices should be explicitly justified and carefully aligned with the underlying conception of fairness. To support this process, we provide a framework of criteria and practical guidelines for ethically grounded fairness assessment, with the aim of fostering more transparent and accountable evaluation of intersectional bias in AI systems.

CCS Concepts: • **General and reference** → **Metrics**; **Reliability**; • **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**.

Additional Key Words and Phrases: Intersectionality, Meta-Metric, Bias, Aggregation, Fairness, Theoretical Alignment, Assessment, Accountability

ACM Reference Format:
Jeanne Monnier, Thomas George, Frédéric Guyard, Christèle Tarnec, and Marios Kountouris. 2026. Beyond Base Metric: The Critical and Overlooked Role of Meta-Metrics in Intersectional Fairness. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3805689.3812417

1 Introduction

Previous studies have shown that machine learning (ML) systems can not only reproduce but also amplify societal biases encoded in data [34]. As these systems are increasingly deployed in high-stakes domains, questions of fairness have become central to their evaluation and governance. Fairness in AI/ML is commonly conceptualized through individual, group, and causal approaches [11]. We focus on *group fairness*, which models the population

Authors' Contact Information: Jeanne Monnier, EURECOM, Sophia Antipolis, France and Orange Research, Sophia Antipolis, France, jeanne.monnier@orange.com; Thomas George, Orange Research, Châtillon, France, thomas.george@orange.com; Frédéric Guyard, Orange Research, Sophia Antipolis, France; Christèle Tarnec, Orange Research, Sophia Antipolis, France; Marios Kountouris, EURECOM, Sophia Antipolis, France and University of Granada, Granada, Spain, mariosk@ugr.es.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812417

as a collection of demographic groups defined by shared sensitive attributes, and seeks to prevent unjustified disparities across groups.

Fairness is a complex concept embedded in a rich socio-technical context, and capturing it may require situated, in-depth analyses grounded in a detailed understanding of the application domain. However, such approaches are not always compatible with broad practical applicability. In practice, the assessment, comparison, and mitigation of unfairness in ML systems often rely on fairness metrics that reduce complex normative considerations to a single summary score, much like performance metrics. For example, in-processing mitigation methods typically incorporate fairness through a constraint or regularization term, which must be expressed as a mathematically tractable summary quantity that can be included in the learning objective.

In this paper, we make a technical contribution to improving the tools used to assess intersectional fairness in ML. We argue that existing tools are not always sufficiently expressive or conceptually grounded to ensure alignment with a defensible conception of fairness. Although our contribution is formal and mathematical, it has a cross-disciplinary scope: it seeks to bridge in-depth social science analyses of fairness and the practical need to operationalize fairness in mathematical terms. To this end, we broaden the space of evaluation choices and provide a framework to help practitioners select metrics that are better aligned with their normative commitments and application context.

The choice of metrics plays a crucial role in shaping how complex social phenomena are interpreted and acted upon. By compressing rich, multidimensional realities into interpretable summaries, metrics inevitably embed normative and methodological assumptions. This challenge is especially acute in the context of fairness, an essentially contested ethical concept whose meaning depends on normative commitments and contextual factors [22]. Fairness metrics must therefore be aligned with the underlying conception of fairness they are intended to operationalize. The literature has extensively examined different conceptions of fairness, such as treating similar individuals similarly or ensuring statistical parity across demographic groups. Even within group fairness, disagreement persists over which outcomes should be equalized, e.g., positive rates, true positive rates, or error rates, leading to a proliferation of often incompatible fairness definitions [33]. Moreover, any given definition may itself admit multiple interpretations [2]. Although these alternatives may yield different conclusions, each may be legitimate depending on the context. Metric choices must therefore be carefully aligned with the underlying conception of fairness.

Most works on group fairness assume a binary setting in which the population is partitioned into two groups along a single sensitive attribute, such as gender. Although analytically convenient, this framing is often too simplistic. In practice, discrimination is *intersectional*: multiple social attributes interact to shape individuals' experiences and outcomes. Sociological research has shown that intersectionality is non-additive [9], such that the effects associated with an intersectional identity may exceed the sum of the effects associated with its individual components. Analyzing sensitive attributes independently can therefore obscure important patterns of disadvantage, which makes fairness evaluation in intersectional settings essential.

This paper focuses on a previously underexplored stage of fairness measurement at which additional normative ambiguity enters: the *aggregation of intersectional subgroup-level measures into a single summary fairness score*. This aggregation is already performed implicitly whenever machine learning evaluation moves beyond binary settings, though it has received little explicit attention in the literature, likely because intersectional fairness itself remains comparatively underexamined. Because it necessarily involves more than two demographic groups, intersectionality produces high-dimensional outputs when model behavior is analyzed across groups, and this complexity grows with the number of groups considered. Prior work has sometimes reported such results through tables or visualizations in order to preserve granularity. Although informative, these representations are difficult to interpret directly and are often ill-suited to decision-making contexts (Fig. 1 in [31]). In many real-world settings, however, a single fairness score is needed for tasks such as model selection, auditing, or policy compliance, which are often embedded in highly automated workflows with limited contextual review.

Indeed, in current practice, even in simpler binary-fairness settings, a brief reflection to compare fairness notions is rarely undertaken. Although there is certainly scope for a debate on these practices, it has been observed that the adoption of fairness tools is challenging, despite their simplified and limited scope. It is clear that the extra complexity involved in considering the various ways of approaching aggregation within intersectionality will rarely be rigorously addressed if left to individual practitioners. This paper aims to pragmatically improve rigor within existing workflows to help facilitate and increase accurate use of fairness tools.

Constructing such a score in an intersectional setting requires reducing a multidimensional vector of subgroup disparities to a scalar summary. What is relatively straightforward in binary settings, often captured by a simple difference or ratio [33], admits many alternatives in the intersectional case, thereby introducing an additional layer of measurement modeling. Lum et al. [31] refer to this second layer as a *meta-metric*, distinguishing it from the underlying *base metric* [33]. Yet the choice of meta-metric is often treated as a default or overlooked altogether. Across much of the existing work on intersectional fairness, we observe that the meta-metrics in use largely follow the same paradigm: the worst subgroup-level disparity. As we show, however, a broader range of possibilities exists, and different meta-metrics can yield substantially different, and sometimes conflicting, fairness assessments. Ignoring this design choice risks misalignment between the quantity being measured and the underlying normative conception of fairness.

In this work, we provide a systematic analysis of fairness *meta-metrics*, characterizing their design space along three axes: (1) a *group-comparison step*, which determines how subgroup disparities are defined; (2) a *score-aggregation step*, which combines these disparities into a single summary value; and (3) an *optional weighting step*, which allows different disparities to be emphasized differently. We identify and formalize the principal modeling choices at each stage, and show that different combinations of these choices, and thus different meta-metrics, can lead to markedly different fairness evaluations. Although all meta-metrics coincide under conditions of complete fairness with respect to the chosen base metric, they diverge in the more realistic cases in which disparities persist. In such cases, disparities may be distributed unevenly across subgroups, and each meta-metric implicitly prioritizes different distributions of disadvantage. We analyze the normative implications of these design choices and conclude by offering practical guidelines to help both technical and non-technical stakeholders navigate this moral and methodological challenge whenever a scalar summary is operationally required. More specifically, we provide guidance for selecting meta-metrics that are transparent, contextually appropriate, and ethically aligned.

2 Related Work

Fairness has emerged as an important issue in machine learning, with highly publicized controversies surrounding the use of historically discriminatory data in applications such as COMPAS [1], UCI Adult [3], and German Credit [20]. Technical approaches to fairness in machine learning [11, 34] are commonly categorized into individual fairness [16], causal fairness [30], and group fairness [33] – our main focus.

Most work on group fairness focuses on binary group settings that divide the population into privileged and unprivileged groups. Even work acknowledging intersectionality often reduces to a binary setting when presenting contributions or reporting experiments [13]. This simplification has facilitated understanding, supported the definition of foundational criteria [18], contributed to a comprehensive body of work [6, 11, 33, 35], and enabled the development of mitigation methods [5, 15, 21, 23, 28, 29, 32, 40]. It has also led to practical tools such as IBM's AI Fairness 360 [4], Microsoft's FairLearn [7], and Google's TensorFlow Fairness Indicators, broadening adoption. However, it also risks suggesting that a model can be unbiased by achieving good scores on a single binary sensitive attribute, while allowing substantial bias, including intersectional bias, to persist.

Buolamwini and Gebru [10] highlighted early that binary frameworks oversimplify real-world discrimination, where attributes are plural and multi-valued. The term *subgroup* is commonly used in intersectional contexts to emphasize finer-grained groups defined by intersections of larger groups [10, 39]. Such binary framings can

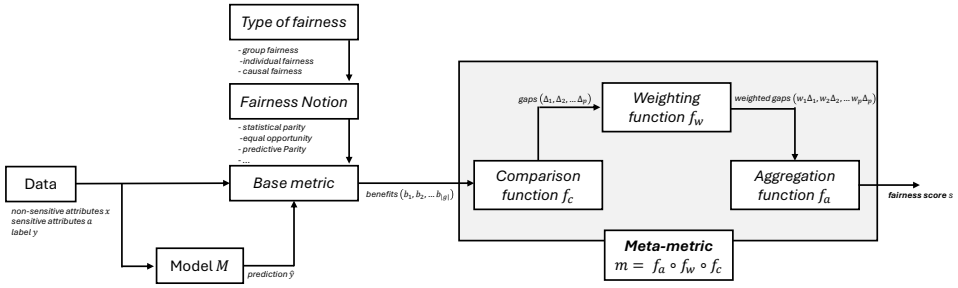


Fig. 1. Complete process of intersectional fairness score building. *After identifying sensitive attributes, constituting demographic subgroups, selecting the appropriate group fairness notion and applying the corresponding base metric, $|G|$ measures of benefits corresponding to each subgroup must be aggregated into a single fairness score s through a meta-metric.*

overlook issues such as fairness gerrymandering [26, 36], related to Simpson's paradox [8], where a particular grouping can mask biases that become apparent when data are subdivided into smaller subgroups. This aligns with the non-additivity described by Bright et al. [9], whereby biases across distinct sensitive attributes do not add up to the bias on their combination.

While intersectional fairness has been surveyed [12, 17], the issue of *meta-metrics* is largely absent. A recurring theme is reliance on worst-case disparity measures [26, 37], which we argue are only one choice among many. For example, Chen et al. [12] define multi-attribute fairness as the "maximum disparity between any two subgroups," without considering alternative aggregation strategies. While such measures are intuitive, they may not always be the most appropriate: minimizing worst-case disparity may be desirable in settings like healthcare, whereas other contexts may prioritize broader parity across subgroups. The lack of discussion of aggregation choices, alongside the absence of alternatives, suggests this approach is treated as default.

We argue that practitioners should leverage available options to align the theoretical conception of fairness with the evaluation metric. In line with Jacobs and Wallach [22] warning against mismatches between constructs and their operationalization when measurement modeling offers multiple choices, we study the related question of making an informed choice of the appropriate *meta-metric*.

The concept of *meta-metrics*—aggregating subgroup disparities into a single fairness score—remains underexplored. Lum et al. [31] introduced the term, contrasting it with *base metrics* that measure performance within groups. Their focus is primarily on the statistical accuracy of bias estimation when subgroup samples are small, a challenge also addressed in [36]. By contrast, the question of which *meta-metric* best captures intersectional fairness has received little attention, despite many plausible aggregation strategies that can yield different assessments. Existing approaches predominantly rely on worst-case measures, which may not always be sufficient. We aim to address this gap by systematically analyzing the space of possible *meta-metrics*, proposing a taxonomy, and emphasizing alignment between fairness theory and measurement practice in intersectional settings.

Independent of our work, [19] identify moral-methodological *desiderata* required for a metric to correctly assess fairness, that is, properties intended to guide the search for appropriate metrics in intersectional settings. Similarly, we provide guidelines to help AI model builders identify the *meta-metric* that aligns with their conception of intersectional fairness. Unlike *desiderata*, which articulate general principles concerning the desirable properties of fairness metrics, our guidelines are intended to facilitate the selection of an appropriate metric for a specific application context. These guidelines therefore require prior reflection on the expectations and normative commitments underlying intersectional fairness in the given setting.

3 A Taxonomy of Meta-Metrics

3.1 Problem Setting

Datasets and fairness. Let D be a dataset of instances $(\mathbf{X}, \mathbf{A}, Y) = (x_1, \ldots, x_m, a_{m+1}, \ldots, a_n, y)$, where each instance corresponds to an individual. The feature vector is composed of non-sensitive attributes $\mathbf{X}$ and sensitive attributes $\mathbf{A}$. y is the label that the model will predict. We denote the sensitive attribute space by $\mathcal{A} = \prod_{i=m+1}^{n} \mathcal{A}_i$, where each $\mathcal{A}_i$ represents the domain of a sensitive attribute for which discrimination is ethically or legally impermissible. The designation of an attribute as sensitive is assumed to result from prior normative and societal deliberation. We denote by $|\mathcal{A}|$ the number of sensitive attributes under consideration. When $|\mathcal{A}| \geq 2$, intersectionality arises. Based on the sensitive attributes, we define a set of *demographic subgroups*, denoted by $g \in G$, where each subgroup consists of individuals sharing identical values across all sensitive attributes. The total number of demographic subgroups considered is denoted by $|G|$. We distinguish two main settings: the binary case, where $\left|G\right| = 2$, which is the focus of most existing fairness literature; and the non-binary case, where $\left|G\right| > 2$. The latter occurs either when intersectionality is present or when a single sensitive attribute takes more than two values.

Base metric. For a classifier M learned using a machine learning algorithm, group fairness is typically evaluated by selecting a *base metric* corresponding to a previously specified *fairness notion.* A *base metric* is a performance measure computed separately for each demographic subgroup. It takes $M(x) = \hat{y}$ the prediction of the model and sometimes y the initial label as inputs to build a new quantity b, which we call the *benefit.* Applying the base metric results in a vector of benefits $(b_1, b_2, ..., b_{|G|})$. For example, statistical parity is commonly assessed via the probability of receiving the advantageous outcome, while overall accuracy equality is assessed via the probability of being correctly classified. More generally, there exist as many base metrics as theoretical conceptions of what fairness in a model should entail [33]. Although our analysis is general and can be extended to different fairness notions, for clarity of exposition, we focus on equal opportunity for illustration, which uses the true positive rate as its base metric. Equal opportunity fairness metrics in use are defined for binary classification, reflecting the historical focus of fairness research on binary decision tasks and datasets. To isolate and examine issues arising specifically from the choice of meta-metric, we adopt the same binary classification setting. While this restriction warrants critical attention, it remains orthogonal to the challenges posed by intersectionality, which persist even under binary outcomes.

Binary case. In the standard binary setting with two demographic groups, the final fairness metric is obtained by directly comparing the values of the base metric computed for each group. For example, under equal opportunity, comparing the true positive rates, this yields the following measure: $\mathrm{EO}_M = |P(\hat{y} = 1 | y = 1, a = 1) - P(\hat{y} = 1 | y = 1, a = 0)|$, which quantifies the disparity in the likelihood of receiving the positive outcome $\hat{y} = 1$, conditional on the true label $y = 1$, between the two groups defined by the protected attribute a. While alternative formulations, such as using ratios instead of differences, are possible, the modeling of fairness metrics in the binary case remains relatively straightforward and results directly in a single scalar fairness score. Consequently, when comparing two models M_1 and M_2, one can determine which model is fairer by simply comparing their respective fairness scores (scalars).

Intersectionality: the need to summarize multiple measurements. Intersectionality gives rise to settings in which the number of demographic subgroups satisfies $|G| > 2$. Applying a base metric in this context yields a "benefit score" for each subgroup. Comparing these scores across all pairs of subgroups results in $\frac{|G|(|G|-1)}{2}$ pairwise disparity measurements, or gaps $\Delta_{i,j}$, a quantity that grows rapidly with the number of sensitive attributes and their possible values. While an exhaustive analysis of these pairwise disparities can provide a detailed view of a model's potential biases, such representations are not synthetic and are difficult to interpret, compare, or

operationalize. This tension motivates the need for principled methods to aggregate multiple subgroup-level measurements into a coherent summary.

Meta-Metrics. In the pursuit of a single scalar value to represent intersectional fairness, the apparent simplicity of the binary case is often replicated without explicit justification. However, while the choice between differences and ratios in binary settings is relatively limited and has modest consequences, aggregating multiple subgroup-level disparities in intersectional contexts admits many distinct modeling choices, which can lead to non-comparable and even conflicting results. Despite its importance, this aggregation step has received little explicit attention in the fairness literature, thereby jeopardizing the validity and interpretability of resulting fairness assessments. The choice of aggregation method is consequential: subtle differences in how subgroup disparities are combined can reverse fairness rankings between models. That is, a model M_1 may be judged fairer than a model M_2 under one aggregation strategy, yet less fair under another. This variability underscores the need to systematically identify available aggregation choices and to understand their normative and practical implications. In the following sections, we introduce a taxonomy of fairness meta-metrics derived from an analysis of existing aggregation practices. This taxonomy makes explicit the key decision points involved in constructing a meta-metric and clarifies the range of available options at each stage, providing a foundation for more transparent and ethically aligned fairness measurement in intersectional settings.

Section 3.2 studies the first modeling choice: given the values of the base metric, how should the 'benefit' received by each subgroup be compared to that received by others? Section 3.3 then addresses the second design choice: once each subgroup's benefit has been situated relative to the others by measuring gaps, how should these relative values be aggregated into a single scalar reflecting the overall fairness of the system? Section 3.4 explores an additional degree of freedom, namely the possibility of assigning weights to subgroups or to pairs of subgroups in order to differentially emphasize certain disparities. Taken together, these three axes define a family of meta-metrics, each of which is the composite function of these three components. We summarize this design space in Table 1.

3.2 First modeling choice: comparison type

As a first step, the base metric is applied separately for each demographic subgroup. This produces a $\mathbb{R}^{|G|}$ vector of subgroup-level performance, the benefit vector, typically expressed as probabilities conditioned on group membership.

In this work, we are not primarily concerned with the absolute values of these subgroup-level measurements, but rather with their relative relationships. Fairness, in this context, is understood as the absence of unjustified disparities between groups. Accordingly, we introduce a first function, denoted f_c, which takes as input a vector of size $|G|$ and compares each subgroup's performance to that of the others. This comparison step constitutes the first key modeling choice in the construction of a meta-metric. It determines how disparities are identified and characterized, and thus whether, and in what sense, bias is deemed to be present. It results in a new vector of gaps Δ. We identify and analyze two distinct comparison strategies, which we introduce below.

In the **one-vs-all** approach (e.g., [27]), the value measured for each subgroup is compared individually against the values measured for all other subgroups. This results in a collection of pairwise gaps across all distinct subgroup pairs, arranged as a vector of disparities $\Delta_{i,j}$ corresponding to the $\frac{|G|(|G|-1)}{2}$ unique unordered pairs (i, j) of subgroups.

$$\Delta^{\text{EO}}_{\text{o-v-all}}(M, i, j) := P(\hat{y} = 1 | y = 1, g = g_i) - P(\hat{y} = 1 | y = 1, g = g_j)$$

Thus, we obtain a vector $\Delta^{\text{EO}}_{\text{o-v-all}}(M) = \left(\Delta^{\text{EO}}_{\text{o-v-all}}(M, i, j)\right)_{(i,j)}$ of size $\frac{|G|(|G|-1)}{2}$.

In the **one-vs-mean** approach (e.g., [13]), the value measured for each subgroup is compared to the mean performance of all subgroups. This yields a vector of $|G|$ values, each representing the difference between a subgroup's benefits and the average benefit across the population.

$$\Delta^{\text{EO}}_{\text{o-v-mean}}(M, i) := P(\hat{y} = 1 | y = 1, g = g_i) - P(\hat{y} = 1 | y = 1)$$

Thus, we obtain a vector $\Delta^{\text{EO}}_{\text{o-v-mean}}(M) = \left(\Delta^{\text{EO}}_{\text{o-v-mean}}(M, i)\right)_i$ of size $|G|$.

In addition to the choice of comparison design, a further modeling decision concerns the comparison operation itself: in both approaches described above, subgroup performances may be compared using ratios instead of differences. In the binary setting, this distinction constituted the only available degree of freedom in fairness metric design. In intersectional settings, however, combining the possible comparison designs with the choice of comparison operation yields four distinct comparison variants at this first modeling stage.

3.3 Second modeling choice: aggregation method

The comparison stage results in structured information about disparities between subgroups, but a single fairness score has yet to be determined. In this respect, the situation mirrors that of binary fairness metrics prior to their final scalar formulation. The challenge is no longer to identify or quantify the presence of bias, but rather to synthesize the multiple disparity measurements into a single indicator reflecting the overall extent of unfairness. As discussed earlier, such aggregation is necessary both for human interpretability and for practical use cases, such as comparing models against one another or evaluating them with respect to a predefined criterion. Consequently, the vector of comparison outcomes—whose dimensionality depends on the chosen comparison type—must be aggregated using a function f_a that maps it to a scalar value in $\mathbb{R}$. Combining one of the comparison strategies introduced in Section 3.2 with one of the aggregation methods presented below yields a complete meta-metric, enabling fairness assessment in intersectional settings.

We identify four major types of mathematical tools that can be used for aggregation: q-norms, ordered weighted averaging (OWA), threshold-based indicators, and independence criteria.

3.3.1 L_q norms. Let $q \in \mathbb{R}^+$. The L_q norm of a vector $(\delta_1, \delta_2, \ldots, \delta_n)$ is defined as

$$\|\delta\|_q = \left(\sum_{i=1}^{n} | \delta_i |^q \right)^{1/q}$$

L_q norms provide a flexible measure of the magnitude of a vector. Larger values of q place increasing emphasis on the largest components of the vector, converging in the limit $q \to \infty$ to $\|\delta\|_\infty := \max_i |\delta_i|$. **For q = 1**: all subgroup disparities contribute equally, yielding an aggregate measure of the overall accumulation of unfairness across subgroups [13]. **For q = 2**: larger disparities are weighted more heavily, causing the metric to emphasize more severely disadvantaged subgroups. **For q $\to \infty$**: the aggregation is dominated by the most disadvantaged subgroup, corresponding to a worst-case notion of fairness.

3.3.2 Ordered Weighted Averaging. OWA aggregation consists of computing a weighted average of a vector after ordering its elements by magnitude. Let $(\delta'_1, \delta'_2, \ldots, \delta'_n)$ denote the values of $(\delta_1, \delta_2, \ldots, \delta_n)$ sorted in increasing order of magnitude, and let $(w_1, w_2, \ldots, w_n)$ be a vector of non-negative weights assigned according to this ranking, with $\sum_{i=1}^{n} w_i = 1$. Each weight w_i thus determines the importance given to the i-th smallest component of the vector.

$$OWA := \sum_{i=1}^{n} w_i \delta'_i$$

This aggregation method allows the sensitivity of the overall fairness score to be tuned with respect to extreme disparities versus the overall distribution of disparities, offering even greater flexibility than L_q norms. The

weight vector $(w_1, w_2, \ldots, w_n)$ acts as a set of hyperparameters that substantially expands the space of possible aggregation behaviors and makes the method highly customizable. In particular, OWA aggregation subsumes several familiar operators as special cases, including the **minimum**, **maximum**, and **mean**.

3.3.3 Threshold-based indicators. Let $\varepsilon \in [0, 1]$. For any vector $(\delta_1, \delta_2, \ldots, \delta_n)$, the *$\varepsilon$-threshold indicator* measures the proportion of components whose value exceeds the threshold ε:

$$\text{ind}^{\varepsilon} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(\delta_i > \varepsilon\right)$$

The hyperparameter ε determines the tolerance level for acceptable disparities. For small values of ε, the indicator prioritizes avoiding any disparity, regardless of its magnitude. For larger values of ε, the focus shifts toward avoiding only large disparities, while smaller gaps are treated as acceptable. This aggregation method emphasizes whether disparities exceed a predefined threshold rather than how large they are. Unlike norm-based or OWA aggregation, it does not account for the magnitude of disparities beyond the threshold. Instead, it captures the frequency with which subgroup disparities are deemed substantively significant. As a result, this approach places greater emphasis on minimizing the occurrence of unfairness rather than on reducing its severity once it arises.

3.3.4 Independence criterion. Disparities in subgroup performance, as measured by a base metric, can alternatively be interpreted as statistical dependence between performance outcomes and subgroup membership. From this perspective, fairness corresponds to a form of independence between these variables. Mutual information [24, 25] provides a principled criterion for quantifying such dependence. For any pair of random variables X and Y, mutual information measures the amount of information shared between them. It is defined as

$$I(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x, y) \log\left(\frac{P_{X,Y}(x, y)}{P_X(x) P_Y(y)}\right)$$

where $P_{X,Y}$ is the joint distribution of X and Y, and P_X and P_Y are their marginal distributions. Unlike the comparison-based aggregation methods introduced above, mutual information operates directly on subgroup-level outcomes and maps them to a scalar dependence measure without an explicit comparison stage. It is therefore best understood as a direct, non-comparison-based meta-metric within the broader design space considered in this paper. In the illustrative case of equal opportunity, this amounts to computing the mutual information $I(g, \hat{y} \mid y = 1)$, which quantifies the dependence between subgroup membership and the model's predictions among positively labeled instances.

3.4 Optional design choice: weighting

Meta-metrics can be further refined through the introduction of group weights, yielding additional variants. While optional, weighting substantially expands the design space, and choosing not to apply weights is itself a meaningful modeling decision. This step defines f_w, the weighting function.

Prior to aggregation, it is therefore possible to assign weights to each comparison outcome, whether associated with individual subgroups or with pairs of subgroups, depending on the selected comparison strategy. Weighting modifies the relative influence of different disparities on the final fairness score and thus carries important normative implications.

3.4.1 Weighting by sample size. One common option is to weight subgroup-level measurements according to sample size. This reduces the influence of small samples on the aggregate score, whose estimates may be noisy or unrepresentative. In this sense, weighting can help limit the effect of outliers arising from statistical uncertainty.

Aggregation type f_a | Comparison type f_c		$\Delta_{\text{one-vs-all}}(M) := \left(b_{g_i} - b_{g_j}\right)_{(i,j)}$	$\Delta_{\text{one }-vs-\text{ mean}}(M) := \left(b_{g_i} - \bar{b}\right)_i$
L_q norms	L_1	$\|\Delta_{\text{o-v-all}}\|_1 = \sum_{(i,j)} \left\|b_{g_i} - b_{g_j}\right\|$	$\|\Delta_{\text{o-v-mean}}\|_1 = \sum_i \left\|b_{g_i} - \bar{b}\right\|$
	L_2	$\|\Delta_{\text{o-v-all}}\|_2 = \sqrt{\sum_{(i,j)} \left(b_{g_i} - b_{g_j}\right)^2}$	$\|\Delta_{\text{o-v-mean}}\|_2 = \sqrt{\sum_i \left(b_{g_i} - \bar{b}\right)^2}$
	L_q	$\|\Delta_{\text{o-v-all}}\|_q = \left(\sum_{(i,j)} \left\|b_{g_i} - b_{g_j}\right\|^q\right)^{1/q}$	$\|\Delta_{\text{o-v-mean}}\|_q = \left(\sum_i \left\|b_{g_i} - \bar{b}\right\|^q\right)^{1/q}$
ind^ε	low ε / high ε	$\text{ind}^\varepsilon_{all} = \frac{2}{n(n-1)} \sum_{(i,j)} \mathbf{1}\left(\left\|b_{g_i} - b_{g_j}\right\| > \varepsilon\right)$	$\text{ind}^\varepsilon_{mean} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(\left\|b_{g_i} - \bar{b}\right\| > \varepsilon\right)$
Ind. Criterion	Mutual Information	$(b_1, b_2, ..., b_{\|G\|})$	
		$\mathrm{I}(g;b) = \sum_{\gamma \in \mathcal{G}} \sum_{\beta \in \mathcal{B}} P_{g,b}(\gamma,\beta) \log \frac{P_{g,b}(\gamma,\beta)}{P_g(\gamma) P_b(\beta)}$	

Table 1. **Comprehensive taxonomy of meta-metrics, presented in Section 3**. For the sake of simplicity, weighted variants have not been added. Each meta-metric of this table is a composite function of a comparison function f_c (see 3.2) and an aggregation function f_a (see Section 3.3). b denotes benefits computed by the selected base metric.

Conversely, fairness analyses often deliberately emphasize small or underrepresented subgroups, which may correspond to populations that are less visible and more vulnerable to discrimination. From this perspective, it may be desirable to assign greater weight to such subgroups, for example, by weighting inversely proportional to sample size. These contrasting choices illustrate how weighting schemes encode normative priorities and must be selected with care.

3.4.2 Weighting by subgroup distances. In the case of a **one-vs-all** comparison, subgroup pairs may differ in the number of sensitive-attribute values they share. Disparities between subgroups that are close in this sense, i.e., that differ on only one sensitive attribute (e.g., young white men vs. young white women), may be viewed as less severe than disparities across multiple attributes (e.g., young white men vs. older non-white women). Accordingly, meta-metrics may weight subgroup-pair disparities as a function of their distance in the sensitive-attribute space, assigning greater weight to disparities between more dissimilar subgroups. Alternatively, particular sensitive attributes may be deemed more salient than others, in which case disparities involving differences along those attributes can be assigned higher weights. These choices allow stakeholders to encode domain-specific or normative priorities into the fairness assessment.

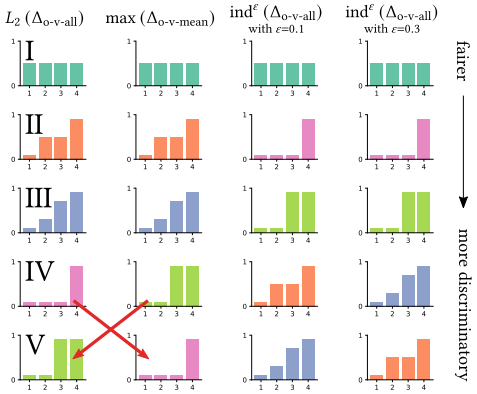


Fig. 2. Rankings induced by different meta-metrics applied to the same subgroup-level values. Different choices of comparison type and aggregation method can produce divergent (red arrows), and even reversed, fairness rankings.

In the end, comparison-based meta-metrics in our taxonomy can be written as the composite function $m = f_a \circ f_w \circ f_c : \mathbb{R}^{|G|} \to \mathbb{R}$, which maps the benefit vector to a scalar fairness score. Direct dependence-based meta-metrics such as mutual information instead operate directly on subgroup-level outcomes without instantiating the comparison stage. Table 1 summarizes the comparison-based meta-metrics in our taxonomy.

4 Empirical Illustration of Meta-Metric Choices

4.1 Different meta-metrics can induce different rankings

Our first research question is whether the meta-metrics introduced above are equivalent, or whether they can induce different rankings over the same set of distributions. To provide an answer, we try to establish the existence of cases in which different meta-metrics rank the same candidate distributions differently.

4.1.1 Illustrative Example on Synthetic Data. If such cases exist, then the issue at stake, namely, the non-equivalence of the meta-metrics introduced above, is substantive and warrants careful examination. To investigate this question, we begin by considering hypothetical subgroup distributions that are not directly derived from empirical data but remain plausibly observable in practice. We examine five distinct hypothetical distribution profiles, each representing subgroup-level values produced by the base metric (Fig. 2).

Each histogram, colored green (I), orange (II), blue (III), pink (IV), or yellow-green (V), depicts outcomes of the base-metric for four demographic subgroups, with each bar corresponding to one subgroup. In the illustrative case of equal opportunity, each bar represents the true positive rate $P(\hat{y} = 1 \mid y = 1, g = g_i)$ for subgroup g_i. We selected these five profiles because they capture some of the most extreme configurations in the space of possible subgroup-outcome distributions. For each profile, we compute a fairness score using several meta-metrics. We then rank the profiles separately under each meta-metric, from the fairest (top) to the most discriminatory (bottom).

The green histogram serves as a consistency check. It corresponds to the idealized case in which all subgroups attain identical base-metric values, implying the absence of disparities and, therefore, no group-level unfairness. This profile represents a widely accepted fairness benchmark. As expected, all considered meta-metrics rank it as the fairest outcome.

The remaining four distribution profiles were selected to illustrate distinct patterns of disparity. They share the same extreme values: in all cases, the lowest base-metric value is 0.1 and the highest is 0.9. Beyond this common feature, however, disparities are distributed across subgroups in substantially different ways. The blue profile exhibits relatively small pairwise disparities across subgroups, yet no two subgroups attain identical base-metric values. By contrast, the pink profile contains three subgroups with identical scores and a single subgroup whose score differs markedly from the others, thereby concentrating a large disparity on one group. These profiles therefore correspond to qualitatively distinct practical situations, a distinction that we discuss further in Section 5.

First, we note that, under a worst-case aggregation rule—specifically, the maximum pairwise disparity, denoted $\max(\Delta_{\text{one-vs-all}})$—all four profiles receive the same fairness score. This result illustrates how certain aggregation choices can obscure substantively meaningful differences in the distribution of subgroup disparities.

Fig. 2 presents the rankings of these same distribution profiles under four meta-metrics. Notably, none of these meta-metrics induces the same ranking. A distribution ranked as fairest under one meta-metric may be ranked as least fair under another. This pattern is illustrated by the orange distribution, which is ranked best by $L_2(\Delta_{\text{one-vs-all}})$ but worst by $\text{ind}^\epsilon(\Delta_{\text{one-vs-all}})$. In the context of bias mitigation, this divergence means using $L_2(\Delta_{\text{one-vs-all}})$ might suggest that the orange distribution is sufficiently fair to justify halting further intervention. By contrast, using $\text{ind}^\epsilon(\Delta_{\text{one-vs-all}})$ would motivate continued mitigation efforts, pushing the system toward a profile closer to the pink distribution, which this meta-metric ranks best after the green distribution. Moreover, for every pair of distribution profiles, there exists at least one meta-metric that ranks the pair in one order and another that ranks them in the opposite order (one example is highlighted by the red arrows in the figure). This observation further confirms that different meta-metrics do not induce a unique or absolute partial order. We discuss the practical consequences of these observations for model assessment, comparison, and bias mitigation in Section 4.1.2.

This first illustrative case demonstrates that meta-metrics are not equivalent: changing the meta-metric can alter the resulting fairness rankings and even yield contradictory conclusions. This result highlights the risks of overlooking this stage of intersectional fairness measurement or adopting a meta-metric by default without explicit justification. Although this existence proof is based on theoretical distribution profiles, these profiles are plausibly observable in practice. The problem identified here is therefore not merely theoretical, as we show in Section 4.1.2.

4.1.2 Real-World Data Application. Having established the existence of this problem through a theoretical example, we now illustrate it using a real-world dataset. Specifically, we train multiple models on the UCI Adult dataset [3] using several standard ML methods: support vector classifiers with radial basis function (RBF), linear, and polynomial (poly) kernels; Decision Trees (dct); Bagging (bag); Random Forests (rf); and Gradient Boosting (boost). All models are trained with the default hyperparameters provided by the scikit-learn library [38], yielding 7 distinct models in total. Once trained, we apply the models to the test set and evaluate subgroup-level values under the fairness notion of equal opportunity using the base metric $|P(\hat{y} = 1 \mid y = 1, g = g_i)|$, which we estimate empirically by counting. In this experiment, we consider 10 subgroups defined by the two sensitive attributes *Sex* and *Race*. We then evaluate and rank the models using different meta-metrics derived from the taxonomy introduced earlier. As in the theoretical analysis, we observe substantially different rankings of the seven models depending on the meta-metric design choices. The resulting rankings are presented in Tables 3a and 3b. In each table, each column reports the ranking of the seven models under the meta-metric indicated in the column heading. We present the results from two complementary perspectives. In Table 3a, the comparison type, when applicable, is fixed to *one-vs-all*, while the aggregation method varies across the options introduced in our taxonomy. By contrast, Table 3b holds the aggregation method constant while varying the comparison type between *one-vs-all* and *one-vs-mean*.

	$L_1(\Delta_{\text{o-v-all}})$	$L_2(\Delta_{\text{o-v-all}})$	$L_\infty(\Delta_{\text{o-v-all}})$	$\text{ind}^{0.2}(\Delta_{all})$	$\text{ind}^{0.3}(\Delta_{all})$	$I(A;B)$
rbf	6	4	1	7	7	1
linear	1	1	2	2	4	6
poly	5	3	2	6	6	2
dct	7	7	6	1	5	7
bag	2	5	4	2	2	3
rf	3	2	7	5	2	5
boost	4	6	5	2	1	4

(a) Rankings of the 7 models under different aggregation methods: L_1, L_2, L_∞, $\text{ind}^{0.2}$, $\text{ind}^{0.3}$, and $I(g,b)$.

	$\text{ind}^{0.2}(\Delta_{all})$	$\text{ind}^{0.2}(\Delta_{mean})$	$\text{ind}^{0.3}(\Delta_{all})$	$\text{ind}^{0.3}(\Delta_{mean})$
rbf	7	6	7	6
linear	2	3	4	1
poly	6	3	6	1
dct	1	6	5	6
bag	2	1	2	4
rf	5	1	2	4
boost	2	3	1	1

(b) Rankings of the 7 models under different comparison methods (*one-vs-all* and *one-vs-mean*)

Fig. 3. Rankings of models trained on the Adult dataset, assessed using multiple meta-metrics from the taxonomy of Section 3.

In both tables, the models identified as the fairest (blue) or most biased (red) are not consistent across meta-metrics. Considered jointly, the two tables reveal that each model is designated as the fairest by at least one meta-metric. This finding confirms that model rankings are highly sensitive to the choice of meta-metric, and that different evaluations can induce conflicting pairwise orderings. Moreover, for several models, one can identify one meta-metric under which the model ranks first and another under which it ranks last. For example, the SVC with RBF kernel is ranked as the fairest model according to mutual information $I(g,b)$ and $L_\infty(\Delta_{\text{o-v-all}})$, yet as the most biased according to all threshold-based indicators. By contrast, the Decision Tree model performs poorly under norm-based evaluations but is ranked as the fairest according to $\text{ind}^{0.2}(\Delta_{all})$.

This practical application allows us to confirm two points. First, evaluating intersectional fairness using the range of meta-metrics introduced above is feasible on a real-world dataset. Second, the differences observed in the theoretical analysis (Section 4.1.1) also arise in practice when these meta-metrics are applied to real-world data. In practical terms, if the seven models had been compared to determine which was most desirable from a

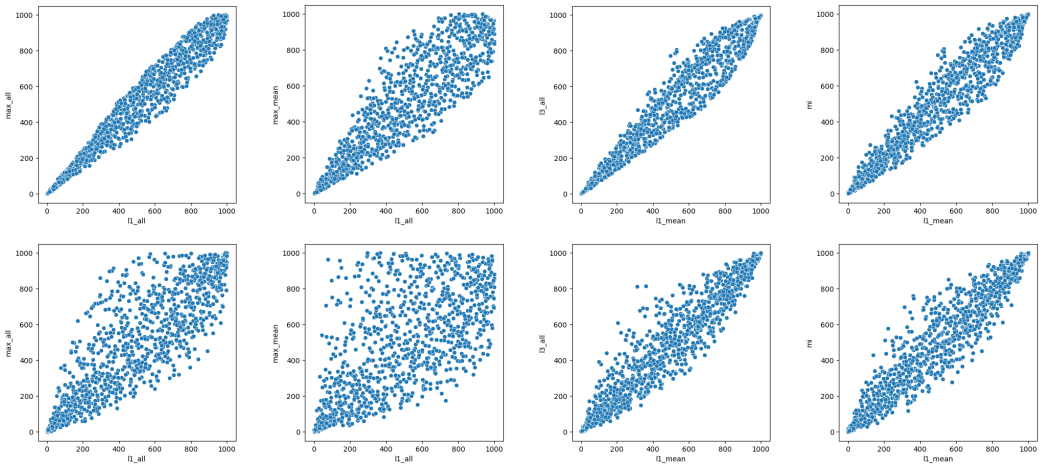


Fig. 4. Pairwise rank correlations between meta-metrics computed over 1000 randomly sampled distributions, for 4 subgroups (**top**) and 8 subgroups (**bottom**)

fairness perspective, the choice of meta-metric could lead to diametrically opposed conclusions. Likewise, if the goal had been to reject models based on bias, some models identified as fairest under one meta-metric would have been rejected under another. Similarly, in the context of bias mitigation, a model deemed acceptable under one meta-metric might instead have been judged to require further intervention under another. Taken together, these results illustrate the practical significance of the problem identified.

4.2 Systematically quantifying pairwise differences between aggregation methods

As a second illustrative case, we quantitatively examine the relationships between different meta-metrics. We randomly sample 1000 distributions of subgroup-level benefits for 4 subgroups and another 1000 distributions for 8 subgroups. For each distribution, we compute fairness scores using various meta-metrics and we then compare the resulting rankings (Figure 4).

For the setting with four subgroups (top row), the resulting rankings are generally strongly correlated, with Kendall's rank correlation in the range of $\tau \in [0.73, 0.87]$. By contrast, increasing the number of subgroups to 8 (bottom row) substantially decreases rank correlations between pairs of meta-metrics, with $\tau \in [0.49, 0.79]$.

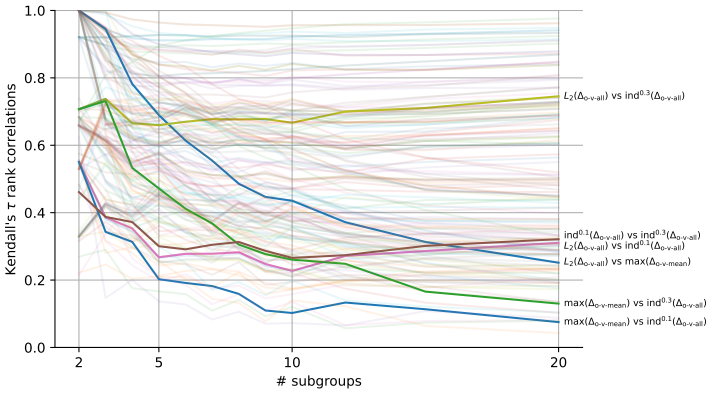


Fig. 5. Correlations between rankings obtained using different meta-metrics, as the number of subgroups grows.

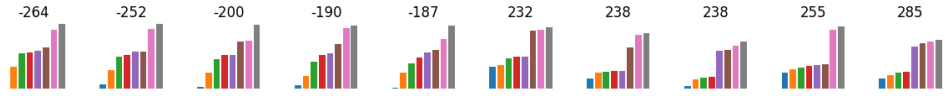
To support these observations, we perform a more systematic study of the effect of a varying number of subgroups (here in the range 2 to 20) on

pairwise correlations in Figure 5. We observe an overall trend of decreasing pairwise correlation as the number of subgroups grows, with some exceptions (see highlighted representative cases).

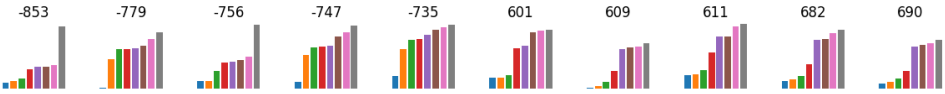
As the number of sensitive attributes increases, the number of intersectional subgroups grows exponentially (e.g. adding a single binary sensitive attribute doubles the number of subgroups), suggesting that divergence between meta-metric rankings is likely to become more pronounced. These results indicate that the choice of meta-metric becomes increasingly consequential as intersectional complexity increases, and should therefore be treated with particular care.

4.3 Qualitative study of meta-metrics

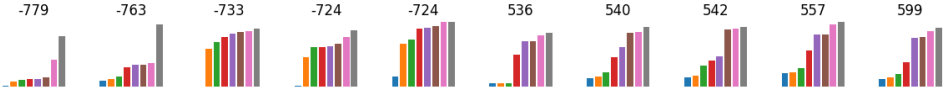
Sections 4.1 and 4.2 demonstrate that meta-metrics are not equivalent, and that some induce rankings that are more closely correlated than others. We now examine selected meta-metrics from a qualitative perspective, with the aim of understanding what kinds of distributional patterns they emphasize or penalize, ultimately to help guide practitioners in their choices. This analysis informs the discussion in the following section on why one meta-metric may be preferable to another in a given context. We begin by noting that distributions receiving the best fairness scores are generally uninformative, as they correspond, regardless of the meta-metric, to highly uniform subgroup outcomes. By contrast, the worst-ranked distributions are more revealing, as they highlight the types of disparity patterns that a given meta-metric penalizes. A complementary source of insight comes from examining distributions whose rankings differ most sharply between two meta-metrics. Such cases expose patterns that one meta-metric treats permissively while another treats strictly, thereby making explicit the differing normative priorities encoded by each aggregation strategy.



(a) Distributions with the largest ranking differences under $L_1(\Delta_{one-vs-mean})$ and $L_1(\Delta_{one-vs-all})$.



(b) Distributions with the largest ranking differences under $L_1(\Delta_{one-vs-mean})$ and $max(\Delta_{one-vs-mean})$.



(c) Distributions with the largest ranking differences under $L_1(\Delta_{one-vs-all})$ and $max(\Delta_{one-vs-mean})$.

Fig. 6. Each figure displays, for a given pair of meta-metrics (m_1, m_2), the distributions that are ranked most differently by the two meta-metrics. Distributions shown on the **left** (resp. **right**) are ranked more favorably by m_1 (resp. m_2), with the corresponding rank difference indicated above.

For a given pair of meta-metrics, we select, from our random sample of 1,000 distributions, the ten distributions with the largest ranking discrepancies: five ranked more favorably by the first meta-metric and five by the second. The resulting distributions are shown in Figure 6. These results provide additional insight into the relative proximity of different meta-metrics. Consistent with the correlation analysis presented earlier, the magnitude of rank differences varies substantially across meta-metric pairs. We focus on the most extreme discrepancies observed. For the pair $L_1(\Delta_{\text{one-vs-mean}})$ and $L_1(\Delta_{\text{one-vs-all}})$ (Figure 6a), the maximum observed difference is 285

ranking positions. In contrast, for the pair $L_1(\Delta_{\text{one-vs-mean}})$ and $\max(\Delta_{\text{one-vs-mean}})$ (Figure 6b), the maximum difference reaches 853 positions, suggesting that these two meta-metrics are substantially less aligned.
With respect to distributional patterns, several consistent preferences emerge. When contrasted with either $L_1(\Delta_{\text{one-vs-all}})$ or $\max(\Delta_{\text{one-vs-mean}})$, the meta-metric $L_1(\Delta_{\text{one-vs-mean}})$ tends to assign poorer rankings to distributions exhibiting clustered subgroup structures. In such patterns, most pairwise disparities $\Delta_{(i,j)}$ remain below a moderate threshold, but a small number of disparities are markedly large. By contrast, both $L_1(\Delta_{\text{one-vs-all}})$ and $\max(\Delta_{\text{one-vs-mean}})$ appear to penalize this type of distribution less severely. This difference reflects the distinct sensitivities of these meta-metrics to the frequency versus the magnitude of subgroup disparities.
Considering Figures 6a and 6c, further observations can be made on $L_1(\Delta_{\text{one-vs-all}})$. In both cases, this meta-metric tends to penalize progressive distributions in which subgroup disparities are consistently small but never equal to zero, corresponding to gradually increasing profiles.
Finally, a recurring pattern among distributions that are strongly penalized by $\max(\Delta_{\text{one-vs-mean}})$ consists of isolated subgroup cases, in which a single subgroup exhibits large disparities with all other subgroups, while the remaining subgroups are tightly clustered in terms of benefit.

5 Guidelines

We now translate these quantitative analyses into practical guidelines to support the selection of an appropriate meta-metric by practitioners for a given context. The central question is how to meaningfully characterize a distribution of disparities across subgroups. Given an appropriately selected fairness notion, any disparity in base-metric performance across subgroups constitutes a form of bias. While the absence of disparities -a perfectly uniform distribution of subgroup-level outcomes- is the unanimous ideal fairness scenario (green distribution in Figure 2), assessing fairness in the situations where disparities persist is inherently ambiguous. Different patterns of disparity may reflect distinct ethical concerns, and their relative severity depends on how fairness is conceptualized.

5.1 Pattern preferences

In Section 4, we identified recurring distributional patterns across subgroup outcomes. We propose to use these patterns as the basis for a framework that supports the interpretation and comparison of different disparity distributions. Reflecting on the implications of these patterns in concrete application contexts helps determine which configurations are ethically acceptable and which should be avoided. Building on the empirical observations presented earlier, we link the various design options and hyperparameters of the taxonomy introduced in Section 3 to their respective sensitivities, positive or negative, to particular distributional patterns. This mapping enables us to articulate practical guidelines for constructing meta-metrics that are aligned with a desired conception of intersectional fairness.

A first recurring pattern is the **isolated subgroup** pattern (pink case in Figure 2). This corresponds to situations in which a large majority of subgroups receive similar outcomes, while one subgroup experiences substantially worse or better treatment. In practical terms, such a pattern reflects a trade-off in which near-ideal fairness is achieved for most subgroups at the expense of severe (dis)advantage for a few. To discourage this type of pattern, meta-metrics such as $\text{ind}^{\varepsilon}(\Delta_{\text{one-vs-all}})$ (with a high value of ε) and $\max(\Delta_{\text{one-vs-mean}})$ are particularly effective. The former penalizes the frequency of large pairwise gaps, while the latter directly penalizes situations in which a subgroup deviates strongly from the average outcome.

A related pattern arises when the population is divided into **two or more blocks of subgroups** (yellow-green case in Figure 2), each block exhibiting internally equal outcomes, but with substantial disparities between blocks. Compared to isolated-subgroup scenarios, the severity of 'individual' disadvantage at subgroup-level may be

reduced, but a larger portion of subgroups experiences unfair treatment. When these blocks are distributed symmetrically, the overall average remains centered. Such patterns are therefore particularly penalized by meta-metrics that emphasize numerous deviations from the mean, such as $L_q(\Delta_{\text{one-vs-mean}})$ with low values of q, or $\text{ind}^{\varepsilon}(\Delta_{\text{one-vs-mean}})$ with low values of ε.

Conversely, some distributions exhibit a more **continuous spread of benefits** (blue case in Figure 2), in which disparities between subgroups are relatively small but consistently present. In such cases, no pair of subgroups receives identical outcomes: differences are pervasive, yet large gaps between any particular pair of subgroups are uncommon. This type of pattern is discouraged by meta-metrics such as $\text{ind}^{\varepsilon}(\Delta_{\text{one-vs-all}})$ with low values of ε, which penalize high frequencies of disparities, even when they are small. It is also penalized by $L_q(\Delta_{\text{one-vs-all}})$ with low values of q, which emphasizes the cumulative magnitude of pairwise differences across subgroups. These choices reflect a preference for eliminating systemic inequality.

In this subsection, we have characterized and linked different patterns of disparity distributions to the meta-metrics that tend to penalize or favor them. Importantly, none of the identified profiles is inherently more problematic than another. As with many ethical questions, the assessment of fairness is inherently context-dependent. When situated within a specific application domain, certain patterns may emerge as more concerning than others, depending on the values, goals, and constraints at stake.

5.2 End-to-End Case studies

We briefly illustrate how meta-metric selection depends on context by applying the process proposed in this paper to two well-known datasets (see Fig. 1). These examples are not intended to defend particular normative assumptions, but to show how different application settings can motivate different meta-metric choices. A full justification of such choices requires domain-specific and social science expertise and lies beyond the scope of this work.

For a criminal justice setting such as COMPAS [1], plausible sensitive attributes include *Sex* and *Race*. If the primary concern is wrongful incarceration, the relevant base metric is *Predictive Equality*, which focuses on false positives [14]. Under this framing, what matters most is avoiding cases in which one subgroup bears a markedly higher risk of unjust deprivation of liberty. This perspective supports attention to one-vs-all disparities and, in particular, to the worst such disparity. A plausible meta-metric is therefore $\text{ind}^{\varepsilon}(\Delta_{\text{one-vs-all}})$ with high ε value.

For the UCI Adult dataset [3], plausible sensitive attributes include *Sex*, *Race*, and *Marital Status*. If the normative objective is to identify broad, persistent inequalities in income outcomes, *Statistical Parity* is a plausible base metric to address historical biases. In this setting, the central concern is less an isolated extreme case than a pervasive spread of unequal benefits across subgroups. This perspective supports a meta-metric such as $L_1(\Delta_{\text{one-vs-all}})$, which penalizes the cumulative amount of subgroup disparities.

6 Conclusion

In this work, we formalized the design of meta-metrics and identified available design choices by establishing a three-axis framework comprising group comparison, score aggregation, and optional weighting. We show that different combinations of these choices can yield markedly different fairness evaluations. Because these choices encode distinct views of which disparity patterns should be prioritized, meta-metric selection should be treated as a substantive design decision that requires explicit justification.

Our contribution is necessarily constrained by its formal and technical scope. Determining which disparity patterns should be prioritized in a given context requires interdisciplinary engagement, particularly with social scientists and domain experts. Nevertheless, we hope this work helps address a largely unexamined practice in intersectional fairness research, the uncritical use of meta-metrics, and provides a foundation for more transparent, rigorous, and contextually grounded fairness assessment.

Generative AI Disclosure Statement

We declare that we did not use any generative AI tools during manuscript preparation.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Richard Arneson. 2018. Four Conceptions of Equal Opportunity. *The Economic Journal* 128, 612 (2018), F152–F173. doi:10.1111/ecoj.12531

[3] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. doi:10.24432/C5XW20

[4] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2019. AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4:1–4:15. doi:10.1147/JRD.2019.2942287

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. arXiv preprint arXiv:1706.02409. arXiv:1706.02409 [stat.ML] https://arxiv.org/abs/1706.02409

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (2021), 3–44. doi:10.1177/0049124118782533

[7] Sarah Bird, Miroslav Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, and Kathleen Walker. 2020. *Fairlearn: A Toolkit for Assessing and Improving Fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[8] Colin R. Blyth. 1972. On Simpson's Paradox and the Sure-Thing Principle. *J. Amer. Statist. Assoc.* 67, 338 (1972), 364–366. doi:10.1080/01621459.1972.10482387

[9] Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. 2016. Causally Interpreting Intersectionality Theory. *Philosophy of Science* 83, 1 (2016), 60–81.

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. PMLR, New York, NY, USA, 77–91.

[11] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A Clarification of the Nuances in the Fairness Metrics Landscape. *Scientific Reports* 12, 1 (2022), 4209. doi:10.1038/s41598-022-07939-1

[12] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2024. Fairness Improvement with Multiple Protected Attributes: How Far Are We?. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. Association for Computing Machinery, New York, NY, USA, 1–13.

[13] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. 2020. A Fair Classifier Using Mutual Information. In *Proceedings of the 2020 IEEE International Symposium on Information Theory*. IEEE, Piscataway, NJ, USA, 2521–2526. doi:10.1109/ISIT44484.2020.9174163

[14] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 797–806. doi:10.1145/3097983.3098095

[15] Hyungrok Do, Preston Putzel, Axel S. Martin, Padhraic Smyth, and Judy Zhong. 2022. Fair Generalized Linear Models with a Convex Penalty. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, PMLR, 5286–5308.

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. doi:10.1145/2090236.2090255

[17] Usman Gohar and Lu Cheng. 2023. A survey on intersectional fairness in machine learning: notions, mitigation, and challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 6619–6627.

[18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc., Red Hook, NY, USA, 3315–3323.

[19] Johannes Himmelreich, Arbie Hsu, Ellen Veomett, and Kristian Lum. 2025. The Intersectionality Problem for Algorithmic Fairness. In *Proceedings of the Algorithmic Fairness Through the Lens of Metrics and Evaluation (Proceedings of Machine Learning Research, Vol. 279)*. PMLR, 68–95.

[20] Hans Hofmann. 1994. German Credit Data. UCI Machine Learning Repository. doi:10.24432/C5NC77

[21] Vasileios Iosifidis and Eirini Ntoutsi. 2019. AdaFair: Cumulative Fairness Adaptive Boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 781–790. doi:10.1145/3357384.3357974

[22] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 375–385. doi:10.1145/3442188.3445901

[23] Heinrich Jiang and Ofir Nachum. 2020. Identifying and Correcting Label Bias in Machine Learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*. PMLR, PMLR, 702–712.

[24] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, Heidelberg, Germany, 35–50. doi:10.1007/978-3-642-33486-3_3

[25] Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. 2022. Infofair: Information-Theoretic Intersectional Fairness. In *Proceedings of the 2022 IEEE International Conference on Big Data*. IEEE, Piscataway, NJ, USA, 1455–1464.

[26] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Stockholm, Sweden, 2564–2572.

[27] Kenji Kobayashi and Yuri Nakao. 2021. One-vs.-One Mitigation of Intersectional Bias: A General Method for Extending Fairness-Aware Binary Classification. In *Proceedings of the International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence*. Springer, Cham, Switzerland, 43–54.

[28] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-Aware Classification. In *Companion Proceedings of the The Web Conference 2018*. International World Wide Web Conferences Steering Committee, Geneva, Switzerland, 853–862.

[29] Peizhao Li and Hongfu Liu. 2022. Achieving Fairness at No Utility Cost via Data Reweighing with Influence. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, PMLR, 12917–12930.

[30] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. 2018. Causal Reasoning for Algorithmic Fairness. arXiv preprint arXiv:1805.05859. arXiv:1805.05859 [cs.LG] https://arxiv.org/abs/1805.05859

[31] Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-Biasing "Bias" Measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 379–389. doi:10.1145/3531146.3533105

[32] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc., Red Hook, NY, USA, 6150–6160.

[33] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. Machine Learning Fairness Notions: Bridging the Gap with Real-World Applications. *Information Processing & Management* 58, 5 (2021), 102642. doi:10.1016/j.ipm.2021.102642

[34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6, Article 115 (2021), 35 pages. doi:10.1145/3457607

[35] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. arXiv preprint arXiv:1811.07867. arXiv:1811.07867 [cs.LG] https://arxiv.org/abs/1811.07867

[36] Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. 2019. Auditing and Achieving Intersectional Fairness in Classification Problems. arXiv preprint arXiv:1911.01468. arXiv:1911.01468 [cs.LG] https://arxiv.org/abs/1911.01468

[37] Jiří Němeček, Mark Kozdoba, Illia Kryvoviaz, Tomáš Pevný, and Jakub Mareček. 2025. Bias Detection via Maximum Subgroup Discrepancy. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 2174–2185. doi:10.1145/3711896.3736857

[38] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[39] Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 336–349. doi:10.1145/3531146.3533101

[40] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 335–340. doi:10.1145/3278721.3278779