# Beyond Consistency: Nuanced Metrics for Individual Fairness

Madeleine Waller
King's College London
London, United Kingdom
madeleine.waller@kcl.ac.uk

Odinaldo Rodrigues
King's College London
London, United Kingdom
odinaldo.rodrigues@kcl.ac.uk

Oana Cocarascu
King's College London
London, United Kingdom
oana.cocarascu@kcl.ac.uk

## Abstract

Individual fairness is the principle aiming for equitable treatment for each individual affected by decisions. Despite its intuitive appeal, the practical applications of individual fairness for algorithmic decision-making systems remain relatively unexplored. In this paper, we investigate the consistency score metric and demonstrate how it fails to adequately capture fairness at the individual level, underscoring the need for a more fine-grained approach. We show that (1) the consistency score obscures instances where individuals are treated significantly differently to the individuals most similar to them and (2) the perceived fairness of individual decisions can be affected by several factors, including the similarity notion itself. To address these issues, we propose four new metrics that measure different aspects of the treatment of individuals with respect to similar individuals, under varying similarity definitions. Our comprehensive evaluation of the new metrics shows that they offer a more nuanced approach to assessing individual fairness, enabling decision-makers to focus on individuals most adversely affected by controversial decisions.

## CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Supervised learning**.

## Keywords

Algorithmic Fairness

**ACM Reference Format:**

Madeleine Waller, Odinaldo Rodrigues, and Oana Cocarascu. 2025. Beyond Consistency: Nuanced Metrics for Individual Fairness. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3715275.3732141

## 1 Introduction

Decision-making processes that have the potential to influence individuals' lives ought to provide guarantees about the fairness of their decisions [36]. This includes decisions made using machine learning (ML), which may carry even greater significance due to the potential scale of their use, impact, and opacity. The field of algorithmic fairness has emerged to address these concerns and aims to achieve equitable outcomes in algorithmic decision-making systems (ADMS) [28]. ADMS are being adopted across a wide range of domains such as criminal justice [27], healthcare [4], finance [11]. However, ADMS can have discriminatory effects, disproportionately impacting certain groups or individuals (e.g. Amazon's hiring tool which was biased against women due to the underrepresentation of previously hired women [10]).

Discrimination is often interpreted and regulated within the framework of legal definitions or societal norms [6, 14, 31, 38]. The field of algorithmic fairness has attempted to map these to mathematical and technical definitions with the aim to quantify unfairness in outputs from ML models [19, 33]. The focus has been on developing fairness metrics for binary ML classifiers, particularly those trained on tabular datasets that contain data about individuals [17] such as protected characteristics (e.g. race, gender, age) [2, 29, 30]. The metrics aim to measure how fair the decisions are for a group or individual. Group fairness requires that different groups receive similar positive and negative classifications across all values of a protected characteristic. Caveats of group fairness metrics have been discussed extensively, with guidance produced on practical scenarios where they are applicable [32].

In contrast, individual fairness is less explored [17]. It focuses on whether the decision for an individual *1)* is similar to decisions for similar individuals or *2)* it would be the same if the individual's protected characteristics were different. There is room for a more in-depth analysis about the nuances of various definitions and alternative notions that aid individuals affected by the decisions made about them.

To this end, we examine individual fairness, focusing on the *consistency score* metric, which is specifically designed for binary classification systems where outcomes correspond to positive or negative decisions [41].[1] The consistency score is a popular individual fairness metric and is also implemented in IBM Fairness 360 [3], one of the few available fairness toolkits, and the only one that incorporates individual fairness metrics [24]. We experiment with the consistency score and varying definitions of similar individuals on three commonly used datasets in the fairness literature and find that it has two main limitations. Firstly, we show that the perceived fairness of individual decisions can be affected by several factors, including the similarity notion itself, the number of similar individuals considered, and the proportion of similar individuals whose decision we wish to agree with an individual's decision. We find that the consistency score does not capture some critical aspects important from an individual fairness perspective, as it averages over all individuals, thus representing an aggregate of individual fairness. Secondly, we demonstrate that the consistency score metric fails to highlight cases where individuals are treated significantly differently to the individuals most similar to them.



*FAccT '25, Athens, Greece*



ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732141

[1] Here we assume that the classifications directly inform decision-making, thus any bias present in the classifications translates to discriminatory decisions.

We propose four novel metrics that overcome the issues with the consistency score. Similarity Robustness of Individual Consistency (*SRIC*) quantifies the overall variation in individual consistency under alternative similarity definitions. Low Individual Consistency Count (*LICC*), Proportional Consistency Score (*PCS*), and Balanced Conditioned Consistency (*BCC*) provide a more nuanced representation of individual fairness where individuals who receive different classifications to the majority of their similar individuals are given greater weight. Our new metrics provide a more fine-grained representation of individual fairness which focuses on actual individuals impacted and accounts for the degree to which decisions for each individual are the same or different to the decisions for similar individuals.

Our **contributions** are as follows: *1)* We show experimentally that the consistency score does not sufficiently quantify individual fairness. *2)* We define the notion of *individual consistency* and propose four individual fairness metrics that measure different aspects of the treatment of individuals with respect to similar individuals, under varying similarity definitions. *3)* We conduct a comprehensive evaluation of the new metrics on three commonly used datasets (Adult Census, COMPAS, German Credit) and show that they offer a more nuanced approach to assessing individual fairness.[2]

## 2 Individual Fairness

Individual fairness aims to ensure individual decisions are non-discriminatory with respect to similar individuals [5]. In the following, we give an overview of metrics for individual fairness, and compare individual and group fairness.

### 2.1 Metrics

There are two[3] main notions of individual fairness: counterfactual and consistency fairness. Various metrics have been proposed to quantify these notions for ADMS based on binary classification (see [1] for a comprehensive background on algorithmic fairness and methods to improve fairness with respect to different metrics).

*2.1.1 Counterfactual.* Counterfactual fairness indicates that *an individual's classification is the same given any value of a protected characteristic* [21]. It is not concerned with individuals similar to the queried individual, rather with hypothetical variations of an individual's attributes. In other words, it implies that the classifications are only impacted by legitimate characteristics and are not affected by an individual's belonging to a protected group [1]. Specifically, counterfactual fairness is satisfied when $P(Y_a = y \mid x, a) = P(Y_{a'} = y \mid x, a')$, where $Y_a$ is the classification for individual $x$ with $a$ as the values of the protected characteristics and $Y_{a'}$ is the classification for an individual with the same attribute values as $x$, except for the values $a'$ of the protected characteristics.

*2.1.2 Consistency.* Consistency dictates that the same decision should be made for individuals with similar characteristics. Various consistency metrics for binary classification have been proposed [12, 41], each requiring a definition of what makes individuals similar. The similarity definition can either be provided by a domain expert given a specific task [12]; found experimentally from the dataset [26]; or use a general $k$ nearest neighbours definition [9, 41].

*Consistency Score.* Our focus in this paper is the *consistency score* [41], which measures the overall level of disparity in the classification of individuals with respect to similar individuals, using the $k$ nearest neighbours (Equation 1).

Let $E$ be a set of unlabelled individuals whose individual fairness of the classifications we want to evaluate, having attributes $Z = \langle z_1, z_2, ..., z_p \rangle$ and domains $\langle D_1, D_2, ..., D_p \rangle$, respectively. Let $v(x, z_i) \in D_i$ be the value of the attribute $z_i$ for individual $x$ and $f : E \rightarrow \{0, 1\}$ a binary classifier for $E$, which we assume to be fixed. Let $nbr_\sigma(x, k) \subseteq E$ be the set of $k$ individuals closest to $x$, according to some notion of similarity $\sigma$. Equation 1 defines $E$'s consistency score [41] with respect to $f$, $\sigma$ and $k$.[4]

$$C^{\sigma,k}(E) = 1 - \frac{1}{|E|} \sum_{x \in E} |f(x) - \frac{1}{k} \sum_{y \in nbr_\sigma(x,k)} f(y)| \qquad (1)$$

$C^{\sigma,k}(E)$ gives the aggregated level of similarity of the classifications of all individuals with respect to their $k$ closest neighbours. It is easy to see from Equation 1 that if all neighbours of an individual receive the same classification as the individual's, the contribution of that individual to the term being subtracted from 1 will be 0. If the classifications of all individuals behave this way, the average of the contributions of all individuals will be 0 and hence the overall consistency score will be 1. Conversely, if an individual's classification differs from all of its neighbours', then the contribution of that individual to the term being subtracted will be $1/|E|$, and if all individuals' classifications behave this way, 1 will be subtracted in total, bringing the consistency score to 0. As a result, $C^{\sigma,k}(E)$ gives a value in the interval $[0, 1]$, and the higher the value, the lower the level of disparity *overall* there is in $E$ with respect to $f$'s classifications. In Section 3.3, we will see that this aggregation potentially masks high levels of disparity for particular individuals.

*Fairness through Awareness.* In contrast to the consistency score, the notion of individual fairness proposed by Dwork et al. [12] requires domain knowledge to define similarity. Specifically, individual fairness is defined as $D(M(x_1), M(x_2)) \leq d(x_1, x_2), \forall x_1, x_2 \in E$ where $M(x)$ is the probability distribution for the output for $x$, $E$ is a dataset, $D$ returns the statistical distance between $M(x_1)$ and $M(x_2)$ and $d$ returns a distance between $x_1$ and $x_2$, found using a pre-specified, domain-dependent definition of similarity. This notion can be used as an optimisation function for model training [12, 39]. Variations include: relaxing the constraint by adding an allowable margin between the distances [39], considering cases where the distance between individuals is small, yet the protected attributes differ [40], and enforcing Fairness through Awareness in ranking decision-making systems as opposed to binary classification systems [22]. We focus on the consistency score as although Fairness through Awareness represents a similar notion, it is not often formalised as a metric for evaluating ADMS, usually embedded into bias mitigation methods as an optimisation function and requires domain knowledge to define $d$.

[2] Code is available at: https://github.com/maddiewaller/MetricsForIndividualFairness

[3] A third notion of individual fairness guarantees that "individuals with greater merit always do better than those with less merit" [13]. However, we will not consider it in this paper as it risks replicating existing biases about who qualifies as having greater merit [20] and is less frequently explored in the algorithmic fairness literature.

[4] The original published paper has an incorrect formulation of consistency which has since been updated by the authors.

**Table 1: Accuracy of the XGBoost classifier and train/test splits for the three datasets. $\%f(x) = 1$ represents the proportion of classifications that are 1. Numerical attributes are shown in italics and unordered categorical attributes are left unformatted.**

| | Train | Test | Accuracy | $\%f(x) = 1$ | Attributes |
|---|---|---|---|---|---|
| Adult | 30162 | 15060 | 0.867 | 21.3% | *age*, workclass, education, marital-status, occupation, relationship, sex, *capital-gain*, *capital-loss*, *hours-per-week* |
| COMPAS | 4933 | 1234 | 0.643 | 64.0% | *age*, sex, race, *juvenile_felony_count*, *priors_count*, *juvenile_misdemeanour_count*, *juvenile_other_count* *charge_degree*, charge_desc |
| German Credit | 800 | 200 | 0.790 | 74.5% | *age*, *checking_status*, *duration*, *purpose*, *credit_amount*, *savings_status*, *installment_commitment*, housing, job, sex |

*Other methods.* Some approaches formalise similarity definitions in a given context based on incomplete knowledge gathered from domain experts, thus being able to find groups of similar individuals in unseen data [18, 37]. The approach by Horesh et al. [16] involves domain experts labelling pairs of individuals that should receive the same classification, in order to measure how closely a classifier's output aligns with these labels. Using a labelled subsample of similar pairs, Lahoti et al. [23] provide a representation of the input data which embeds fairness by ensuring such pairs are treated the same. Exploring how domain experts can assist in defining similarity is an important line of work but outside the scope of our paper.

## 2.2 Individual Fairness vs Group Fairness

There are several key differences between individual and group fairness. Individual fairness does not necessarily require pre-specifying protected characteristics (e.g. consistency metrics focus on ensuring similar treatment of similar individuals, which does not depend on knowledge of the protected characteristics). Further, identical individuals are guaranteed to receive the same decision when individual fairness is satisfied, but this may not necessarily be the case when satisfying group fairness [1]. Individual fairness does not guarantee group fairness except when the distributions of similar individuals for each group are equal [1, 12]. For example, classifying every individual positively satisfies consistency but not some notions of group fairness. This represents an issue as there is no guarantee that a group defined by a protected characteristic will not be treated worse [13]. Finally, individual fairness is arguably a more intuitive notion than group fairness as, in general, it could be worth more to an individual to understand that their decision is fair as opposed to group-wide considerations. For example, showing that similar decisions were made across two groups of a protected characteristic (e.g. the decisions are 80% fair according to some group fairness metric) might not mean much to an individual [35].

# 3 Exploring the Consistency Score

In this section, we delve further into the consistency score metric and assess how it is impacted by varying notions of similarity. To explore the metric, we experiment with three of the most commonly used datasets in the algorithmic fairness literature [17]: Adult Census[5], COMPAS[6], and German Credit[7]. Following Liu et al. [25], we use a reduced number of attributes from each dataset (see Table 1).

[5]https://archive.ics.uci.edu/dataset/2/adult
[6]https://www.kaggle.com/datasets/danofer/compass
[7]https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

We train an XGBoost model on each dataset, using the original training and testing splits for the Adult Census dataset, while randomly selecting 80% of the data for training for the COMPAS and German Credit datasets. Table 1 shows the dataset splits, the model accuracy, the percentage of positive classifications, as well as the attributes used and their type (numerical or unordered categorical). In the remaining of the paper, the metrics and results are reported on the testing sets.

## 3.1 Defining Similar Individuals

The consistency score relies on finding the nearest neighbours of individuals according to some definition of similarity $\sigma$. Defining $\sigma$ depends on several design choices: (1) How to categorise each attribute; (2) How to calculate the distance between the values of each attribute; (3) How to calculate the overall distance between individuals; and (4) How to select similar individuals. We experiment with four definitions of similarity, summarised in Table 2.

We use the attribute values as found in each dataset (see Table 1) for two definitions of similarity, $\sigma_1$ and $\sigma_2$. We also convert the numerical values of attribute 'age' to categorical values (3 categories for $\sigma_3$ and 10 categories for $\sigma_4$), leaving the rest of the attributes unchanged. To compute the distance between attribute values, we use two metrics that differ in the calculation of the numerical attributes: Gower distance (Equation 2) which computes the difference between values normalised for the range of that attribute and Hamming distance (Equation 3) which counts mismatches among attribute values.

$$d_i = \begin{cases} 0, & \text{if } v(e_1, z_i) = v(e_2, z_i) \\ \frac{|v(e_1,z_i) - v(e_2,z_i)|}{\max(D_i) - \min(D_i)}, & \text{if } z_i \text{ is numerical} \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

$$d_i = \begin{cases} 0, & \text{if } v(e_1, z_i) = v(e_2, z_i) \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

We use the Manhattan distance to calculate the overall distance between two individuals: $d(e_1, e_2) = \sum_{i=1}^{p} d_i$. We experiment with several values ($k \in \{5, 10, 15\}$) for the number of similar individuals.

## 3.2 Consistency Score Calculations

Table 3 shows the consistency scores calculated using Equation 1 for the three datasets for different numbers of similar individuals ($k \in \{5, 10, 15\}$) and different notions of similarity ($\sigma_a, a \in \{1, 2, 3, 4\}$)

**Table 2: Different definitions of similarity $\sigma$. For $\sigma_3$ and $\sigma_4$, we convert the numerical values of attribute 'age' to categorical values, leaving the rest of the attributes unchanged.**

| | Attribute categorisation | Distance between attribute values | Overall distance | Selecting number of similar individuals |
|---|---|---|---|---|
| $\sigma_1$ | original | Hamming | Manhattan | $k = 5, 10, 15$ |
| $\sigma_2$ | original | Gower | Manhattan | $k = 5, 10, 15$ |
| $\sigma_3$ | 'age': $< 25, 25-59, \geq 60$ | Gower | Manhattan | $k = 5, 10, 15$ |
| $\sigma_4$ | 'age': $0-10, 11-20, 21-30, 31-40,$ $41-50, 51-60, 61-70, 71-80, 81-90, 90+$ | Gower | Manhattan | $k = 5, 10, 15$ |

**Table 3: Consistency score $C^{\sigma,k}$ for each dataset where $\sigma_a$, $a \in \{1, 2, 3, 4\}$ and $k \in \{5, 10, 15\}$**

| | Adult | | | | COMPAS | | | | German Credit | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ |
| $k = 5$ | 0.894 | 0.912 | 0.900 | 0.905 | 0.677 | 0.716 | 0.710 | 0.713 | 0.689 | 0.725 | 0.708 | 0.707 |
| $k = 10$ | 0.888 | 0.903 | 0.893 | 0.898 | 0.664 | 0.699 | 0.697 | 0.695 | 0.694 | 0.709 | 0.704 | 0.696 |
| $k = 15$ | 0.885 | 0.897 | 0.890 | 0.892 | 0.654 | 0.688 | 0.688 | 0.684 | 0.690 | 0.697 | 0.698 | 0.693 |

as defined in Table 2.[8] The results in Table 3 show that changing the definition of similarity does not greatly impact the consistency score of a dataset. For example, the maximum change in consistency score for the Adult dataset is 0.018, between $\sigma_1$ and $\sigma_2$ when $k = 5$. This small change could be misleading, potentially incorrectly being interpreted as fewer than 2% of individuals have classifications that become more or less 'consistent' with those of similar individuals when the notion of similarity is changed. The change in consistency for the COMPAS and German Credit datasets varies up to 0.04, again showing that changing the definition of similarity does not significantly impact the perceived individual fairness.

The consistency score measures the overall level of disparity in the classifications of a dataset with respect to similar individuals. As the dataset and the individuals' classifications are fixed, it follows that the overall level of disparity stays the same as even if different groups of similar individuals are found, overall the average proportion of classifications that differ in groups is unlikely to differ significantly. The consistency score quantifies the **average** individual fairness and by definition is an aggregate of individual fairness. Thus, it does not sufficiently capture the fairness at an individual level or consider how specific individuals might be impacted. The consistency score provides a generalised view over a whole dataset, but there is no way to assess how consistently **each** individual in the dataset is being treated.

## 3.3 Individual Consistency

In this section we define *individual consistency* based on the individual values (without averaging) in the consistency score in Equation 1 and explore how the consistency score can hide cases in which individuals are assigned different classifications compared to a large proportion of their similar individuals. Recall Equation 1, which computes the consistency score of a dataset $E$. We can see that for a given individual $x \in E$, the term

$$\text{diff}(x) = |f(x) - \frac{1}{k} \sum_{y \in nbr_\sigma(x,k)} f(y)|$$

computes the proportion of $x$'s $k$ most similar neighbours with a classification *different* to $x$'s. For example, if $f(x) = 0$, and for all $y \in nbr_\sigma(x, k)$, $f(y) = 1$, then $\text{diff}(x) = |0 - \frac{k}{k}| = |-1| = 1$. That is, 100% of $x$'s $k$ most similar neighbours have a classification different to it. Given diff, the consistency level of the decision of an individual $x$ is $1 - \text{diff}(x)$.

*Definition 3.1 (Individual consistency score).* Given a binary classifier $f$ for dataset $E$, and an individual $x \in E$, the individual consistency score $c^{\sigma,k}(x)$ computes the proportion of $x$'s $k$ most similar individuals (cf. $\sigma$), with the same classification as $x$'s.

$$c^{\sigma,k}(x) = 1 - |f(x) - \frac{1}{k} \sum_{y \in nbr_\sigma(x,k)} f(y)|$$

Indeed, it is possible to recover $C^{\sigma,k}(E)$ from the individual consistency scores of the individuals in $E$, since

$$\begin{aligned} C^{\sigma,k}(E) &= 1 - \frac{1}{|E|} \sum_{x \in E} |f(x) - \frac{1}{k} \sum_{y \in nbr_\sigma(x,k)} f(y)| \\ &= \frac{1}{|E|} \sum_{x \in E} 1 - |f(x) - \frac{1}{k} \sum_{y \in nbr_\sigma(x,k)} f(y)| \\ &= \frac{1}{|E|} \sum_{x \in E} c^{\sigma,k}(x) \end{aligned}$$

However, we can now use $c^{\sigma,k}(x)$ to look at the decisions of a classifier $f$ at the *individual* level, and to consider the impact of other aspects such as the notion of similarity and the number of most similar neighbours examined. The following definition allows us to identify when an individual has been "consistently" classified by a binary classifier.

[8] In IBM's AI Fairness 360 toolkit [3], the consistency score is calculated using the $knn(x, k)$ function for $nbr_\sigma(x, k)$ which returns the $k$ nearest neighbours of an individual $x$ according to some definition of similarity $\sigma$. $x$ is included in the return value of $knn(x, k)$, thus slightly skewing the score positively. To avoid the oversight in implementation of $nbr_\sigma(x, k)$, we re-implemented the algorithms excluding $x$ from its $k$ nearest neighbours. This slightly reduces the overall consistency values, but does not affect the overall conclusions.

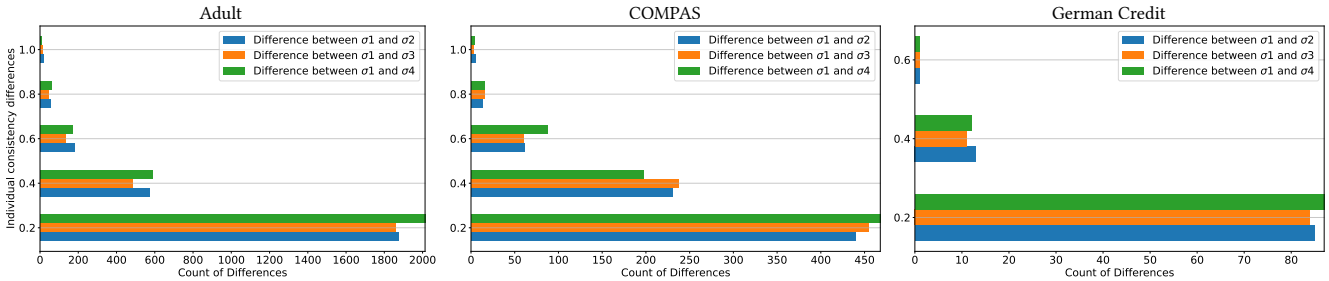


**Figure 1: The differences between individual consistencies across similarity definition $\sigma_1$ and $\sigma_a$, $a \in \{2, 3, 4\}$ with $k = 5$.**

**Table 4: Individual $x$ with ID 9026 (grey) from the Adult dataset and the five similar individuals found using $\sigma_2$ and $\sigma_4$ as in Table 2. The classification $f(x)$ is shown, as well as the distance between each individual and $x$. Here $c^{\sigma_2,5}(x) = 0$ and $c^{\sigma_4,5}(x) = 1$.**

| $\sigma$ | ID | age | workclass | education | marital-status | occ | relationship | sex | capital-gain | capital-loss | hours-per-week | $f(x)$ | Distance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9026 | 26 | *Self-emp-inc* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 32.0 | 0 | 0.0 |
| | 13104 | 38 | *Self-emp-inc* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 50.0 | 1 | 0.348 |
| | 8495 | 43 | *Self-emp-inc* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 44.0 | 1 | 0.355 |
| $\sigma_2$ | 9981 | 44 | *Self-emp-inc* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 45.0 | 1 | 0.379 |
| | 2028 | 47 | *Self-emp-inc* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 45.0 | 1 | 0.42 |
| | 1613 | 47 | *Self-emp-inc* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 45.0 | 1 | 0.42 |
| | 9026 | 21 – 30 | *Self-emp-inc* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 32.0 | 0 | 0.0 |
| | 3335 | 21 – 30 | *Private* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 25.0 | 0 | 1.071 |
| | 7458 | 21 – 30 | *Self-emp-not-inc* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 40.0 | 0 | 1.082 |
| $\sigma_4$ | 4830 | 21 – 30 | *Private* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 40.0 | 0 | 1.082 |
| | 12551 | 21 – 30 | *Private* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 40.0 | 0 | 1.082 |
| | 11033 | 21 – 30 | *Private* | *Some-college* | *Married-civ-spouse* | *Sales* | *Husband* | *Male* | 0.0 | 0.0 | 40.0 | 0 | 1.082 |

*Definition 3.2.* Let $x \in E$ be an individual in the dataset $E$, $\sigma$ a notion of similarity for individuals in $E$, $k$ the number of most similar individuals to consider, and $\delta \in [0, 1]$ a threshold value. Let $f$ be a binary classifier. We say that $x$ has been consistently classified by $f$ (with respect to $x$'s neighbours), if $c^{\sigma,k}(x) \geq \delta$.

This means that the proportion of $x$'s $k$ most similar individuals (cf. $\sigma$) with the same classification as $x$'s (cf. $f$) is at least $\delta$, and indirectly that $f$'s classification for $x$ is acceptable under these specific assumptions. Example 3.3 illustrates how variations in individual consistency score can be masked by the general notion of consistency of a dataset.

*Example 3.3.* Consider the following individual consistency scores for hypothetical dataset $E = \{e_1, e_2, e_3, e_4, e_5\}$, using notions of similarity $\sigma_i$ and $\sigma_j$, and some fixed value $k$.

| | $e_1$ | $e_2$ | $e_3$ | $e_4$ | $e_5$ |
|---|---|---|---|---|---|
| $c^{\sigma_i,k}$ | 0.67 | 0.59 | 0.65 | 0.81 | 0.48 |
| $c^{\sigma_j,k}$ | 0.96 | 0.89 | 0.45 | 0.42 | 0.48 |

The overall consistency score for $E$ is the same ($C^{\sigma_i,k}(E) = C^{\sigma_j,k}(E) = 0.64$), irrespective of the notion of similarity used, as this is the average of the individual consistency scores. Decreases in individual scores can be compensated by increases in others, obscuring the extent to which individuals have classifications different to the majority of their neighbours. In this example, using $\sigma_j$, $e_3$, $e_4$ and $e_5$ have classifications different than the majority of their most similar neighbours – with $\sigma_i$, this only happens with $e_5$. The relatively low scores for $e_3$ and $e_4$ are averaged out by the higher scores for $e_1$ and $e_2$ (cf. $\sigma_j$), and on the surface the consistency of the dataset is the same under both notions of similarity.

*3.3.1 Results & Discussion.* Figure 1 shows the count of the differences in individual consistencies between different definitions of $\sigma$ for Adult, COMPAS, and German Credit datasets. We can see that there are significant changes in individual consistency scores for each dataset, simply by making minor changes in the definition of similarity, as specified in Table 2. These changes in individual consistency are not represented in the consistency score, as seen by the stable values in Table 3. The perceived fairness for an individual can be greatly impacted by changing the definition of similarity, but this will be hidden by the consistency score as it averages over all individuals. To further exemplify this, Table 4 shows an example[9] of an individual in the Adult dataset and the five most similar individuals to it found using $\sigma_2$ and $\sigma_4$, respectively. The individual consistency changes from the minimum of 0, to the maximum of 1, completely changing the perceived fairness for that individual.

## 4 Nuanced Metrics for Individual Fairness

Viewing individual fairness through individual consistency is promising. However, for developers and stakeholders assessing a system as a whole, aggregate measures are also necessary. In this section we define four novel metrics which overcome the limitations discussed of the consistency score not being sufficient to

[9]Chosen arbitrarily from the individuals for which there is a maximum change in individual consistency between two definitions of similarity.

**Table 5: Summary of key features of each metric, detailing their intended evaluation purpose, the type of value produced (P(roportional) or A(bsolute)), and whether they can be used in a loss function.**

| Metric | Purpose | Type | In loss? |
|---|---|---|---|
| *SRIC* | Measures the variation in classification consistency (cf. Definition 3.2) across two notions of similarity. Useful for comparing the impact of switching between two notions. The lower the value, the more resilient the decisions of a classifier $f$ are with respect to the notions. | P | |
| *LICC* | Counts the number of individuals inconsistently classified by $f$ (cf. Definition 3.2). | A | ✓ |
| *PCS* | Measures the proportion of consistent classifications. | P | ✓ |
| *BCC* | Measures the proportion of consistent classifications, taking into account how consistent the decisions are. | P | ✓ |
| *BCC* + penalty | As BCC, but penalises for inconsistent decisions, and hence it is more sensitive to inconsistent decisions. | P | ✓ |

represent individual fairness and use the same experimental setup as in Section 3 to evaluate them.

In Section 3.3 we showed that changing the definition of similarity can change individual consistencies which is not captured by the consistency score. Thus, we propose the Similarity Robustness of Individual Consistency (*SRIC*) metric which quantifies the proportion of individuals for which the individual consistency changes across a threshold between two definitions of similarity.

Our examples in Section 3.3 also highlighted that the consistency score obscures individuals for which the classification differs significantly from the classifications of similar individuals. Thus, we propose metrics that assign greater weight to those individuals with an individual consistency less than a threshold. The Low Individual Consistency Count (*LICC*) quantifies the count of individuals in a dataset for which their individual consistency is below some threshold, highlighting the number of individuals for which we should investigate the individual fairness further. With the same motivation as *LICC*, the Proportional Consistency Score (*PCS*) provides the proportion of individuals for which their individual consistency is above some threshold. To directly compare to the consistency score, we also propose the Balanced Conditioned Consistency (*BCC*) metric which aggregates the individual consistencies, but only for cases where the individual consistencies are above some threshold. An extension of *BCC* incorporates a penalty term to replace any individual consistencies below the threshold, ensuring they are given greater weight in the overall calculation.

Table 5 gives an overview of the key features of each proposed fairness metric to help clarify their intended use cases and guide metric selection based on evaluation goals and constraints.

## 4.1 Similarity Robustness of Individual Consistency (*SRIC*)

The motivation for *SRIC* is to quantify the impact of changing the definition of similarity on the individual consistencies for a dataset. Ideally, the value of *SRIC* should be minimal when small changes are made to the definitions of similarity. This ensures that the perceived fairness is stable to minor adjustments in how we consider or define similarity. Significant variations in *SRIC* may suggest that individuals' perceptions of fairness could differ greatly based on how they define themselves or identify their groups of similar individuals, even when the classifications remain consistent.

*Definition 4.1 (SRIC).* *SRIC* quantifies the proportion of individuals for which the individual consistency *decreases below* or *increases above* some threshold $\delta$, between two different notions of similarity $\sigma_i$ and $\sigma_j$.

$$SRIC_{\delta}^{\{\sigma_i,\sigma_j\},k} = \frac{1}{|E|}\Big|\{x \in E : \big(c^{\sigma_i,k}(x) \geq \delta \wedge c^{\sigma_j,k}(x) < \delta\big) \vee \big(c^{\sigma_i,k}(x) < \delta \wedge c^{\sigma_j,k}(x) \geq \delta\big)\}\Big|$$

*SRIC* is defined in terms of a threshold $\delta$, specifying the proportion of similar individuals with equivalent treatment. For example, $\delta = 0.5$ means there is a shift from the majority of classifications of similar individuals being the same as the classification of the individual $x$ to the minority being the same, or vice versa. Analogously, $\delta = 1$ means that an individual's classification changes from full agreement with the classifications of all similar individuals, to disagreement with the classification of at least one similar individual. In other words, when loss of full agreement occurs.

*4.1.1 Results & Discussion.* Table 6 shows the *SRIC* score with $\delta = 0.5$, the proportion of individuals whose individual consistency changes (Equation 4) and the maximum change in individual consistency (Equation 5) between two definitions of similarity.

$$prop\_change = \frac{|\{x \in E : c^{\sigma_i,k}(x) \neq c^{\sigma_j,k}(x)\}|}{|E|} \tag{4}$$

$$max\_change = \max\left\{c^{\sigma_i,k}(x) - c^{\sigma_j,k}(x)\right\} \tag{5}$$

*SRIC*, *prop_change* and *max_change* are symmetric with respect to $\sigma$, e.g. $SRIC_{\delta}^{\{\sigma_i,\sigma_j\},k} = SRIC_{\delta}^{\{\sigma_j,\sigma_i\},k}$.

Table 6 highlights that although the overall consistency $C_{score}$ stays relatively consistent with changes in the definition of similar individuals as shown in Table 3, the perceived fairness of some individuals is greatly impacted. For example, changing from $\sigma_1$ to $\sigma_2$ in the Adult dataset for $k = 5$, the *SRIC* shows that 5.3% of individual consistencies change from above 0.5 to below 0.5, or vice versa. This demonstrates that a significant proportion of individuals go from being treated the same as the majority of their similar individuals to differently to the majority of their similar individuals, or vice versa. For the same example, *prop_change* shows that 22.8% of individual consistencies change when the definitions of similarity changes and *max_change* shows the maximum variation in individual consistency is 1, meaning at least one individual changes from being

**Table 6: Changes in individual consistency ($c$) between the definitions of similarities in Table 2, with $\delta = 0.5$. Each cell contains *SRIC*, *prop_change*, and *max_change*, in this order.**

| | | Adult | | | | | | | | | COMPAS | | | | | | | | | German Credit | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | $\sigma$ | $\sigma_2$ | | | $\sigma_3$ | | | $\sigma_4$ | | | $\sigma_2$ | | | $\sigma_3$ | | | $\sigma_4$ | | | $\sigma_2$ | | | $\sigma_3$ | | | $\sigma_4$ | | |
| 5 | $\sigma_1$ | 0.053 | 22.8 | 1.00 | 0.057 | 22.1 | 1.00 | 0.057 | 23.3 | 1.00 | 0.209 | 60.7 | 1.00 | 0.236 | 62.4 | 1.00 | 0.231 | 62.6 | 1.00 | 0.140 | 49.5 | 0.60 | 0.140 | 48.0 | 0.60 | 0.135 | 50.0 | 0.60 |
| | $\sigma_2$ | | - | | 0.042 | 16.4 | 1.00 | 0.038 | 16.7 | 1.00 | | - | | 0.143 | 47.7 | 1.00 | 0.156 | 46.8 | 1.00 | | - | | 0.143 | 47.7 | 0.40 | 0.156 | 46.8 | 0.60 |
| | $\sigma_3$ | | - | | | - | | 0.056 | 20.9 | 1.00 | | - | | | - | | 0.204 | 57.4 | 1.00 | | - | | | - | | 0.075 | 39.0 | 0.60 |
| 10 | $\sigma_1$ | 0.035 | 31.6 | 0.90 | 0.043 | 30.8 | 0.90 | 0.042 | 32.3 | 0.90 | 0.165 | 75.0 | 0.80 | 0.202 | 75.5 | 0.80 | 0.186 | 76.9 | 0.80 | 0.065 | 60.5 | 0.40 | 0.070 | 55.5 | 0.40 | 0.065 | 63.5 | 0.40 |
| | $\sigma_2$ | | - | | 0.032 | 23.8 | 0.80 | 0.030 | 26.2 | 1.00 | | - | | 0.135 | 65.0 | 0.80 | 0.125 | 67.1 | 0.80 | | - | | 0.045 | 42.5 | 0.20 | 0.040 | 51.5 | 0.30 |
| | $\sigma_3$ | | - | | | - | | 0.044 | 30.2 | 1.00 | | - | | | - | | 0.181 | 72.6 | 0.90 | | - | | | - | | 0.065 | 63.0 | 0.30 |
| 15 | $\sigma_1$ | 0.036 | 37.4 | 0.80 | 0.042 | 36.6 | 0.87 | 0.042 | 38.6 | 0.80 | 0.192 | 80.5 | 0.73 | 0.218 | 80.7 | 0.73 | 0.179 | 81.0 | 0.73 | 0.075 | 69.5 | 0.27 | 0.090 | 72.0 | 0.33 | 0.085 | 71.0 | 0.33 |
| | $\sigma_2$ | | - | | 0.033 | 27.6 | 0.87 | 0.032 | 32.2 | 0.80 | | - | | 0.159 | 71.2 | 0.67 | 0.146 | 75.4 | 0.60 | | - | | 0.025 | 51.5 | 0.20 | 0.040 | 61.0 | 0.27 |
| | $\sigma_3$ | | - | | | - | | 0.044 | 35.5 | 0.87 | | - | | | - | | 0.211 | 80.5 | 0.80 | | - | | | - | | 0.055 | 64.0 | 0.33 |

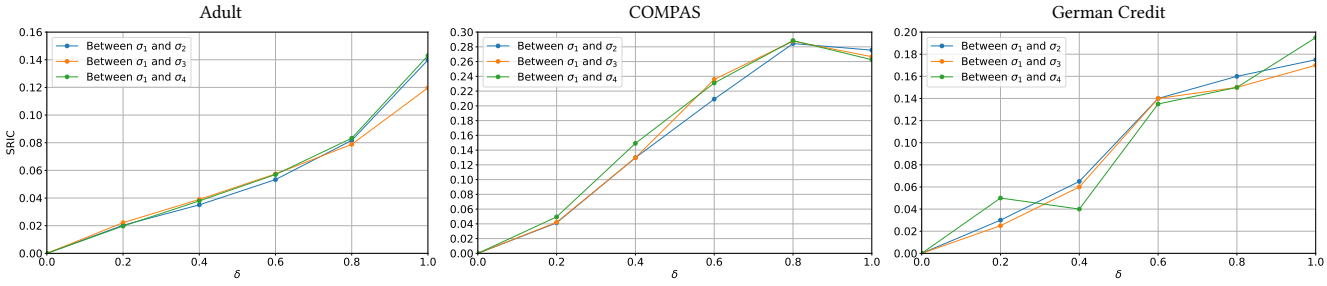


**Figure 2: *SRIC* values between definitions $\sigma_1$ and $\sigma_j$ ($j = 2, 3, 4$), $k = 5$, and values of $\delta$ from 0 to 1 in 0.2 increments.**

classified entirely differently to all of their similar individuals to being classified the same as all of their similar individuals, or the other way around. This occurs often in the Adult and COMPAS datasets. Further, Figure 2 shows how *SRIC* varies for different values of $\delta$ in the three datasets, with $k = 5$, and between definitions $\sigma_1$ and $\sigma_2$ where $\delta$ increases in steps of 0.2 in the interval $[0, 1]$.

## 4.2 Low Individual Consistency Count (*LICC*)

The *LICC* metric measures the absolute number of individuals whose individual consistency score is below some threshold $\delta$. We are arguably more concerned with individuals that have a smaller individual consistency, i.e., individuals for which the classifications are less consistent with the classification of their similar individuals. When aggregating the individual fairness scores to find the overall individual fairness of classifications, individuals with a smaller individual fairness score are hidden. We propose *LICC* as a nuanced quantification of individual fairness that highlights how many individuals are treated significantly different from their similar individuals.

*Definition 4.2 (LICC).* Given a notion of similarity $\sigma$, *LICC* provides a count of individuals for which their individual consistency score is less than some threshold $\delta$.

$$LICC_{\delta}^{\sigma,k}(E) = |\{x \in E : c^{\sigma,k}(x) < \delta\}|$$

*4.2.1 Results & Discussion.* Table 7 shows *LICC* for similarity notions $\sigma_1$–$\sigma_4$ and $\delta = 0.5$. For the Adult dataset, $LICC_{0.5}^{\sigma_1,5}(A) = 1272$ which means that 1272 individuals in the dataset have individual consistency below 0.5, i.e., their classifications differ from the *majority* of the classifications of their 5 most similar individuals. To improve individual fairness, such individuals should be considered carefully due to the risks of them having received unfair treatment.

Figure 3 shows how *LICC* varies for the Adult dataset for all definitions of similarity, where $k = 5$ and $\delta$ increases in steps of 0.2 in the interval $[0, 1]$. When $k = 5$, the *LICC* values for $\delta = 0.5$ and $\delta = 0.6$ are identical and match those in Table 7, as the individual consistencies increase in intervals of 0.2. When $\delta = 1$, the *LICC* score quantifies the number of individuals in the dataset which have at least one similar individual having a different classification. For example, $LICC_1^{\sigma_1,5}(A) = 3559$. This means that 3559 individuals in the Adult dataset (23.6%) have at least one similar individual (cf. $\sigma_1$) with a different classification. These individuals are the ones of particular concern when it comes to mitigating the unfairness in decisions. By using the absolute value, we can also evaluate the scale of impact by finding how many individuals are affected. In contrast to finding a proportion, *LICC* focuses on the individuals themselves, as it could be argued that unfair decisions impacting larger number of individuals carries greater significance, even when the proportion of all decisions remains constant.

## 4.3 Proportional Consistency Score (*PCS*)

*LICC* provides an *absolute* measure of the total number of individuals in a dataset affected by controversial decisions. We can also measure the *proportion* of individuals in the datasets whose decisions are deemed 'consistent', i.e., with an individual consistency score above a certain threshold (cf. Definition 3.2).

*Definition 4.3 (PCS).* Given a similarity notion $\sigma$, the number of similar individuals to consider $k$, and a threshold $\delta$, $PCS_{\delta}^{\sigma,k}(E)$ quantifies the proportion of individuals in $E$ with an individual

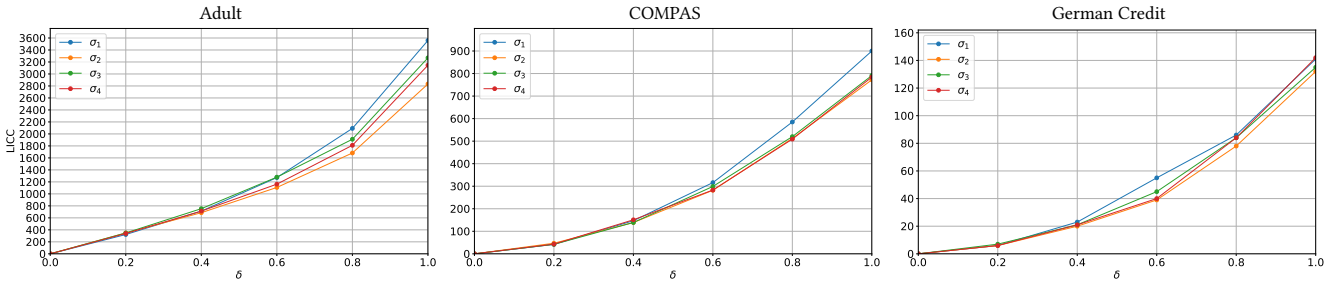


**Figure 3: *LICC* values for similarity definitions $\sigma_1$–$\sigma_4$, $k = 5$, and values of $\delta$ from 0 to 1 in 0.2 increments.**

**Table 7: $LICC_{\delta}^{\sigma_i,k}$ scores where $i \in \{1, 2, 3, 4\}$, $k \in \{5, 10, 15\}$, and $\delta = 0.5$.**

| | Adult | | | | COMPAS | | | | German Credit | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ |
| $k = 5$ | 1272 | 1107 | 1279 | 1162 | 316 | 282 | 299 | 283 | 55 | 39 | 45 | 40 |
| $k = 10$ | 1479 | 1260 | 1421 | 1305 | 399 | 316 | 354 | 345 | 56 | 46 | 51 | 51 |
| $k = 15$ | 1296 | 1157 | 1271 | 1174 | 359 | 284 | 302 | 296 | 48 | 41 | 44 | 45 |

**Table 8: *PCS*, *BCC* and *BCC* with penalty -1, $\delta = 0.5$.**

| | | Adult | | | | COMPAS | | | | German Credit | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ |
| *PCS* | $k = 5$ | 0.916 | 0.926 | 0.915 | 0.923 | 0.744 | 0.771 | 0.758 | 0.771 | 0.725 | 0.805 | 0.775 | 0.8 |
| | $k = 10$ | 0.926 | 0.933 | 0.927 | 0.932 | 0.79 | 0.813 | 0.804 | 0.811 | 0.83 | 0.815 | 0.8 | 0.815 |
| | $k = 15$ | 0.914 | 0.923 | 0.916 | 0.922 | 0.709 | 0.77 | 0.755 | 0.76 | 0.76 | 0.795 | 0.78 | 0.775 |
| *BCC* | $k = 5$ | 0.874 | 0.896 | 0.88 | 0.888 | 0.606 | 0.655 | 0.642 | 0.653 | 0.608 | 0.673 | 0.646 | 0.654 |
| | $k = 10$ | 0.872 | 0.89 | 0.877 | 0.883 | 0.79 | 0.813 | 0.804 | 0.811 | 0.606 | 0.648 | 0.641 | 0.642 |
| | $k = 15$ | 0.862 | 0.879 | 0.868 | 0.872 | 0.555 | 0.61 | 0.606 | 0.604 | 0.608 | 0.635 | 0.632 | 0.621 |
| *BCC* | $k = 5$ | 0.79 | 0.822 | 0.795 | 0.811 | 0.35 | 0.426 | 0.4 | 0.424 | 0.333 | 0.478 | 0.421 | 0.454 |
| +penalty | $k = 10$ | 0.797 | 0.823 | 0.804 | 0.815 | 0.396 | 0.461 | 0.445 | 0.453 | 0.481 | 0.472 | 0.445 | 0.463 |
| | $k = 15$ | 0.776 | 0.802 | 0.784 | 0.795 | 0.265 | 0.38 | 0.362 | 0.364 | 0.368 | 0.43 | 0.412 | 0.396 |

consistency score greater than or equal to $\delta$.

$$PCS_{\delta}^{\sigma,k}(E) = \frac{|\{x \in E : c^{\sigma,k}(x) \geq \delta\}|}{|E|}$$

$LICC_{\delta}^{\sigma,k}(E)$ and $PCS_{\delta}^{\sigma,k}(E)$ are closely related, but $LICC_{\delta}^{\sigma,k}(E)$ gives the absolute number of controversial decisions in $E$, while $PCS_{\delta}^{\sigma,k}(E)$ gives the proportion of acceptable decisions.

$|E| - LICC_{\delta}^{\sigma,k}(E) = |\{x \in E : c^{\sigma,k}(x) \geq \delta\}|$. Hence,

$$PCS_{\delta}^{\sigma,k}(E) = \frac{|E| - LICC_{\delta}^{\sigma,k}(E)}{|E|} = 1 - \frac{LICC_{\delta}^{\sigma,k}(E)}{|E|}.$$

*4.3.1 Results & Discussion.* Recall Example 3.3 where a set of five individuals had the same consistency score (0.64) under two different similarity notions $\sigma_i$ and $\sigma_j$. We saw that under $\sigma_i$ only one individual received a classification different from the majority of the classifications for their similar individuals ($e_5 = 0.48 < 0.5$), whereas under $\sigma_j$ three individuals had individual consistency score *below* 0.5 ($e_3$, $e_4$ and $e_5$). The consistency score is the same under $\sigma_i$ and $\sigma_j$ due to the high individual consistency scores of $e_1$ and $e_2$ under $\sigma_j$ compensating for the low scores of $e_3$ and $e_4$. Using the threshold $\delta = 0.5$, we obtain $PCS_{0.5}^{\sigma_i,k}(E) = 0.80$ (80% of the decisions for individuals match the decisions of the majority of their most similar neighbours), whereas $PCS_{0.5}^{\sigma_j,k}(E) = 0.40$, indicating that only 40% of the decisions for individuals do. This more accurately quantifies the proportion of controversial decisions in the two datasets (see Table 9).

We can compare the values of *PCS* for the datasets in Table 8 with the original values of the consistency score in Table 3. We argue that *PCS* provides a more refined measurement of consistency in the decisions as it emphasises the overall number of individuals whose decisions are questionable. Similarly to *LICC*, the choice of $\delta$ corresponds to the level of individual consistency we find 'acceptable' in a given context. For *PCS*, it is inverse to *LICC* in that $\delta = 1$ requires all of the classifications of similar individuals to be the same as a given individual.

## 4.4 Balanced Conditioned Consistency (*BCC*)

*PCS* measures the proportion of acceptable decisions (i.e., above the threshold $\delta$), but not how *close* to full agreement those decisions were. It allows each individual to contribute to the score with $1/|E|$ as long as their individual consistency score meets the threshold, regardless of how close to the threshold the score is (or how far

from full agreement, i.e., 1, the score is). The balanced conditioned consistency score *BCC* allows us to capture this.

*Definition 4.4 (BCC).* *BCC* gives the sum of the individual consistency scores above or equal some threshold $\delta$ divided by the total number of individuals.

$$BCC_{\delta}^{\sigma,k}(E) = \frac{1}{|E|}\sum_{x\in E} v(x)$$
$$\text{where} \quad v(x) = \begin{cases} c^{\sigma,k}(x), & \text{if } c^{\sigma,k}(x) \geq \delta \\ 0, & \text{otherwise} \end{cases}$$

For fixed parameters, $BCC_{\delta}^{\sigma,k}(E) \leq PCS_{\delta}^{\sigma,k}(E)$.

*4.4.1 Variations of BCC.* We have seen that *BCC* is obtained from a sum of values, where each individual $x$ meeting the consistency threshold $\delta$ contributes with the value $c^{\delta,k}(x)$, or 0 otherwise. This total is then normalised by the total number of individuals. Since $0 \leq c^{\delta,k}(x) \leq 1$, the maximum contribution of each individual is $1/|E|$, and the minimum is 0. It is possible to replace 0 with a suitable *penalty*, effectively bringing down the overall *BCC* score for each controversial decision made. To mirror the positive contribution of each individual, we suggest that the penalty is a value in $[-1, 0]$, and recommend setting it to $-1$. Hence, the maximum penalty for each decision is $-1/|E|$.

*Definition 4.5 (BCC with penalty).* *BCC* with penalty modifies *BCC* by setting a penalty $p \in [-1, 0]$ for each individual consistency score below $\delta$.

$$BCC_{\delta,p}^{\sigma,k}(E) = \frac{1}{|E|}\sum_{x\in E} v(x)$$
$$\text{where} \quad v(x) = \begin{cases} c^{\sigma,k}(x), & \text{if } c^{\sigma,k}(x) \geq \delta \\ p, & \text{otherwise} \end{cases}$$

$BCC_{\delta,-1}^{\sigma,k}(E) \in [-1, 1]$. Notice that having the majority of individuals not meeting the consistency threshold is a sufficient condition for *BCC* with penalty $-1$ to return a negative value, but not a necessary condition. Further, $BCC_{\delta,-1}^{\sigma,k}(E) = -1$ indicates that no individual in $E$ has individual consistency score $\delta$ or greater.

*4.4.2 Results & Discussion.* Consider the consistency scores in Example 3.3 with similarity definitions $\sigma_i$ and $\sigma_j$. $BCC_{0.5}^{\sigma_i,k}(E) = 0.544$, whereas $BCC_{0.5}^{\sigma_j,k}(E) = 0.370$. The lower score using $\sigma_j$ correctly reflects that more individuals ($e_3$ and $e_4$) do not meet the threshold $\delta = 0.5$ than when $\sigma_i$ is used. This shows that *BCC* can be more sensitive to variations in classifications between an individual and their similar individuals, thus better capturing *individual* fairness. Using the same threshold $\delta = 0.5$, but using *BCC* with a penalty of $-1$, $BCC_{0.5,-1}^{\sigma_i,k}(E) = 0.344$, and $BCC_{0.5,-1}^{\sigma_j,k}(E) = -0.230$. *BCC* further discounts individual consistencies which fall below the specified threshold resulting in a sharper distinction between individuals with a high consistency score and those with a low. The closer to $-1$ the penalty is, the more *BCC* amplifies the impact of individual consistencies below the threshold. Table 9 summarises the differences in calculations of *PCS*, *BCC* and *BCC* + penalty, for the set of individuals in Example 3.3, using similarity notions $\sigma_i$ and $\sigma_j$.

**Table 9: *PCS*, *BCC* and *BCC* + penalty scores, for the set *E* of individuals in Example 3.3, using $\delta = 0.5$.**

| | Consistency score | *PCS* | *BCC* | *BCC* + penalty |
|---|---|---|---|---|
| $\sigma_i$ | 0.64 | 0.80 | 0.544 | 0.344 |
| $\sigma_j$ | 0.64 | 0.40 | 0.370 | −0.230 |

Table 8 shows the *BCC* and *BCC* + penalty $-1$ scores for the three datasets. The results show that *BCC* correctly weighs smaller individual consistencies higher and thus lowers the score. For example, for the Adult dataset, $\sigma_1$ with $k = 5$, $C^{\sigma_1,5}(A) = 0.894$, $BCC_{0.5}^{\sigma_1,5} = 0.874$ and $BCC_{0.5,-1}^{\sigma_1,5} = 0.79$. These illustrate that *BCC* (especially with penalty) better reflects the prevalence of controversial decisions in the dataset than *C* does, whilst still giving an indication about the overall level of consistency. We would argue that this offers a more balanced measurement of consistency conditioned by a given threshold of acceptance (and optional penalty). Further, on the COMPAS and German Credit datasets, the variation of *BCC* and *BCC* + penalty scores between definitions of similarity is significantly larger than for the consistency scores. For example, the maximum variation for the consistency score on the German Credit dataset was 0.04. In contrast, the maximum variation for *BCC* with penalty $-1$ is 0.14. This supports our findings that changing the definition of similarity used can change individual consistency scores and thus simply averaging over the individual consistencies hides individuals who are treated differently to a large proportion of their similar individuals.

## 5 Discussion

*Individual Fairness in Law.* The legal significance of individual fairness is rarely discussed due to limited case law involving ADMS [15]. Under certain legal jurisdictions, it is often the responsibility of individuals to raise cases of potential discrimination. For example, in the UK and EU, individuals need to demonstrate that: "a provision, criterion or practice significantly disadvantages a protected group when compared with other people in a similar situation" [33]. In cases involving ADMS, this entails comparing the decision for an individual with those made for similar individuals. The lack of case law makes it challenging to establish clear guidance on how fairness metrics align with existing legal frameworks [32]. Addressing this gap could provide critical insights for both legal and technical communities, ensuring fairness approaches are more robust, interpretable, and actionable within legal contexts. A potential avenue for investigation is whether the decisions for similar individuals, e.g. via consistency metrics, can be used to prove that discrimination occurred. However, individuals do not have access to the whole data, thus the burden of proof would need to shift to the decision-maker to dispel suspicions of unfair treatment. This has the same issues mentioned about which similarity notion to use, how many similar individuals to consider, and so forth.

*Defining Similarity.* For individual fairness, specifically the notion of consistency, we require a definition of similar individuals. In this paper, we focus on the consistency metric as we are able to explore it independently of any domain-specific definition. Existing approaches [16, 18, 22, 37, 39–41] focus on the domain specificity of the metric definitions, requiring input from experts as to how to

define similarity. However, finding suitable domain experts, ensuring their own biases are not transferred, and deciding how to best categorise individuals for the task at hand remain a challenge and limits the use of metrics in the consistency family [7]. The 'ground truth' definition which is often assumed, in reality is rare. Thus, it is important to focus on using metrics that correspond to defined legal definitions of fairness [33, 35]. Further, we assumed a fixed set of $k$ similar individuals. This presents some issues and extensions of our work should consider only individuals that are similar "enough", perhaps defining the set of most similar individuals in terms of a similarity distance threshold. Here, domain expertise is critical to ensure a fair balance between the requirements.

*Exploring Explainability.* Seeing different interpretations of individual fairness for ADMS can improve stakeholders' understanding of a system and help show and prevent individuals being discriminated against. Metrics can be useful for understanding the fairness of a system but in the case of individual fairness, there might be better ways of trying to understand how the system works. Due to the context-dependent nature of fairness in ADMS, metrics alone are not enough. In addition to focusing on socio-technical approaches, we advocate exploring explainability methods to enhance our understanding of individual fairness [34]. Understanding *why* individuals have been treated differently to individuals similar to them is arguably more important than how many were. Explaining how the decision about a particular individual was made and comparing it to the decisions of similar individuals can not only improve the understanding of the system's behaviour (promoting fairness) but also provide re-assurance to the individual about the basis on which the decision was made (promoting transparency and trustworthiness). Counterfactual fairness, for example, is already represented as a field in explainability called counterfactual explainable AI [8] which aims to give causal reasons as to why the classification would change. Similar ideas should be explored for consistency-based individual fairness metrics where different combinations of similar individuals and their classifications should be presented to a domain expert. These would allow the expert to examine, given the context and domain knowledge, whether the categorisations, groupings, and differences in classifications are justified or not, highlighting any cases in which individuals could be discriminated against. Our proposed metrics help highlight the individuals to investigate potential discrimination for. This should be extended to understand *why* there is that disparity, and whether a decision is justifiable in that context.

*Limitations & Future Work.* As previously stated, employing an adequate notion of similarity is use-case dependent and can be challenging. Further, we need to determine at which point neighbours should be considered too dissimilar to be included in the comparison of individual results. As we showed, the evaluations depend on the notion of similarity and this raises the potential risk of gaming the system to achieve misleadingly inflated fairness results.

*LICC*, *PCS*, *BCC*, and *BCC* with penalty can be integrated into a model's loss function depending on the optimisation goal. Penalising for small values of *PCS* (or large values of *LICC*) would maximise the proportion of individuals with individual consistency above $\delta$ while penalising for small values of *BCC* and *BCC* with penalty would maximise the average consistency of individuals with individual consistency above $\delta$. Integrating our metrics into the loss function may involve classification trade-offs, e.g., with accuracy. Since our metrics would optimise for individual fairness, group fairness could also be impacted. For these reasons, at this stage we only use these metrics for detection, leaving the investigation of their use for mitigation for future work. *SRIC* measures the difference in individual consistencies between two definitions of similarity, thus cannot easily be used in a loss function. However, it can be used to assess the resilience of the decisions of a classifier when there are alternative notions of similarity.

Although we only considered examples involving tabular data, our metrics only depend on the ability to compare the decisions of a data point with respect to the decisions of a fixed number of other "close" data points (i.e., the neighbours). Therefore, they are applicable to any dataset where binary classification results are available and a pairwise notion of distance/similarity between data points can be defined, e.g., for text and image data. The investigation of appropriate notions of similarity/distance for specific non-tabular data types is left as future work.

Further, we plan to create a toolkit that allows stakeholders to explore existing metrics and our new metrics, helping to identify if ADMS can be fairly deployed across various domains.

## 6 Conclusion

In this paper, we investigated the application of the consistency score [41], a metric that has been commonly used to quantify the individual fairness of binary classification systems. We demonstrated that the consistency score is not sufficient to faithfully capture important aspects of fairness as it fails to represent fairness at the individual level, particularly in situations where individuals are treated significantly differently from their most similar individuals. Our findings also revealed that altering the definition of similar individuals can significantly affect the proportion of similar individuals with a differing classification to an individual's, ultimately affecting the perceived individual fairness. To address these issues, we proposed four new metrics for individual fairness. Similarity Robustness of Individual Consistency (*SRIC*) quantifies the impact of changing the definition of similar individuals. Low Individual Consistency Count (*LICC*), Proportional Consistency Score (*PCS*), and Balanced Conditioned Consistency (*BCC*) provide new aggregates for individual fairness which give greater consideration to individuals who are treated significantly differently to their similar individuals. Our comprehensive evaluation showed that our metrics provide a more nuanced understanding of individual fairness for binary classification and offer valuable insights for model evaluation and decision-making processes.

## Acknowledgments

This work was supported by the UK Research and Innovation Centre for Doctoral Training in Safe and Trusted Artificial Intelligence[10] [grant number EP/S023356/1]. Madeleine Waller, Odinaldo Rodrigues and Oana Cocarascu are affiliates of the King's Institute for Artificial Intelligence.[11]

[10] https://www.safeandtrustedai.org
[11] https://www.kcl.ac.uk/ai

## References


[1] Joshua W. Anderson and Shyam Visweswaran. 2024. Algorithmic Individual Fairness and Healthcare: A Scoping Review. *medRxiv* (2024). https://doi.org/10.1101/2024.03.25.24304853

[2] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943

[3] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *CoRR* abs/1810.01943 (2018). arXiv:1810.01943 http://arxiv.org/abs/1810.01943

[4] Magnus Bergquist and Bertil Rolandsson. 2022. Exploring ADM in clinical decision-making: Healthcare experts encountering digital automation. In *Everyday Automation*. Routledge, 140–153.

[5] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *FAT* '20: Conference on Fairness, Accountability, and Transparency*. 514–524. https://doi.org/10.1145/3351095.3372864

[6] William Blanchard. 1986. Evaluating Social Equity: What Does Fairness Mean and Can We Measure It? *Policy Studies Journal* 15, 1 (1986). https://www.proquest.com/scholarly-journals/evaluating-social-equity-what-does-fairness-mean/docview/1300125838/se-2

[7] 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.

[8] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim A. Jorge. 2022. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Inf. Fusion* 81 (2022), 59–83. https://doi.org/10.1016/J.INFFUS.2021.11.003

[9] T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27. https://doi.org/10.1109/TIT.1967.1053964

[10] Jeffrey Dastin. 2022. Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women. In *Ethics of Data and Analytics*. Auerbach Publications.

[11] Wendy De La Rosa and Christopher J Bechler. 2024. Unveiling the adverse effects of artificial intelligence on financial decisions via the AI-IMPACT model. *Current Opinion in Psychology* (2024), 101843.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science*. 214–226. https://doi.org/10.1145/2090236.2090255

[13] Will Fleisher. 2021. What's Fair about Individual Fairness?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. 480–490. https://doi.org/10.1145/3461702.3462621

[14] Salvatore Greco, Ke Zhou, Licia Capra, Tania Cerquitelli, and Daniele Quercia. 2024. NLPGuard: A Framework for Mitigating the Use of Protected Attributes by NLP Classifiers. *Proc. ACM Hum. Comput. Interact.* 8, CSCW2 (2024), 1–25. https://doi.org/10.1145/3686924

[15] Ljupcho Grozdanovski. 2021. In search of effectiveness and fairness in proving algorithmic discrimination in EU law. *Common Market Law Review* 58, 1 (2021).

[16] Yair Horesh, Noa Haas, Elhanan Mishraky, Yehezkel S. Resheff, and Shir Meir Lador. 2019. Paired-Consistency: An Example-Based Model-Agnostic Approach to Fairness Regularization in Machine Learning. In *Machine Learning and Knowledge Discovery in Databases - International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part I (Communications in Computer and Information Science, Vol. 1167)*, Peggy Cellier and Kurt Driessens (Eds.). Springer, 590–604. https://doi.org/10.1007/978-3-030-43823-4_47

[17] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2024. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. 1, 2, Article 11 (June 2024), 52 pages. https://doi.org/10.1145/3631326

[18] Christina Ilvento. 2020. Metric Learning for Individual Fairness. In *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020, Harvard University, Cambridge, MA, USA (virtual conference) (LIPIcs, Vol. 156)*, Aaron Roth (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2:1–2:11. https://doi.org/10.4230/LIPICS.FORC.2020.2

[19] Mackenzie Jorgensen, Madeleine Waller, Oana Cocarascu, Natalia Criado, Odinaldo Rodrigues, Jose Such, and Elizabeth Black. 2023. Investigating the Legality of Bias Mitigation Methods in the United Kingdom. *IEEE Technology and Society Magazine* 42, 4 (2023), 87–94. https://doi.org/10.1109/MTS.2023.3341465

[20] Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems*. 325–333. https://proceedings.neurips.cc/paper/2016/hash/eb163727917cbba1eea208541a643e74-Abstract.html

[21] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*. 4066–4076. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[22] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*. IEEE, 1334–1345. https://doi.org/10.1109/ICDE.2019.00121

[23] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. Operationalizing Individual Fairness with Pairwise Fair Representations. *Proceedings of the VLDB Endowment* 13, 4 (2019), 506–518. https://doi.org/10.14778/3372716.3372723

[24] Michelle Seng Ah Lee and Jatinder Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 699:1–699:13. https://doi.org/10.1145/3411764.3445261

[25] Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. 2024. Confronting LLMs with Traditional ML: Rethinking the Fairness of Large Language Models in Tabular Classifications. In *Proceedings of NAACL 2024: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, June 16-21, 2024*, Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (Eds.). ACL, 3603–3620. https://doi.org/10.18653/V1/2024.NAACL-LONG.198

[26] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. 2020. Two Simple Ways to Learn Individual Fairness Metrics from Data. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 7097–7107. https://proceedings.mlr.press/v119/mukherjee20a.html

[27] Northpointe. 2019. Practitioner's Guide to COMPAS Core. https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf

[28] Dana Pessach and Erez Shmueli. 2023. *Algorithmic Fairness*. 867–886. https://doi.org/10.1007/978-3-031-24628-9_37

[29] The United States Department of Justice. 2015. The Fair Housing Act. https://www.justice.gov/crt/fair-housing-act-1

[30] UK Public General Acts. 2010. Equality Act 2010. https://www.legislation.gov.uk/ukpga/2010/15/contents

[31] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness (FairWare '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[32] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *W. Va. L. Rev.* 123 (2020), 735. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772

[33] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law Security Review* 41 (2021), 105567. https://doi.org/10.1016/j.clsr.2021.105567

[34] Madeleine Waller, Odinaldo Rodrigues, and Oana Cocarascu. 2024. Identifying Reasons for Bias: An Argumentation-Based Approach. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 19 (2024), 21664–21672. https://doi.org/10.1609/aaai.v38i19.30165

[35] Madeleine Waller, Odinaldo Rodrigues, Michelle Seng Ah Lee, and Oana Cocarascu. 2024. Bias Mitigation Methods: Applicability, Legality, and Recommendations for Development. *Journal of Artificial Intelligence Research* 81 (2024), 1043–1078.

[36] Madeleine Waller and Paul Waller. 2020. Why predictive algorithms are so risky for public sector bodies. *Available at SSRN 3716166* (2020). https://tinyurl.com/SoRisky

[37] Hanchen Wang, Nina Grgic-Hlaca, Preethi Lahoti, Krishna P. Gummadi, and Adrian Weller. 2019. An Empirical Study on Learning Fairness Metrics for COMPAS Data with Human Supervision. *CoRR* abs/1910.10255 (2019). arXiv:1910.10255 http://arxiv.org/abs/1910.10255

[38] Alice Xiang and Inioluwa Deborah Raji. 2019. On the Legal Compatibility of Fairness Definitions. arXiv:1912.00761

[39] Gal Yona and Guy Rothblum. 2018. Probably approximately metric-fair learning. In *International Conference on Machine Learning*. PMLR, 5680–5688.

[40] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. 2020. Training individually fair ML models with sensitive subspace robustness. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=B1gdkxHFDH

[41] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML*, Vol. 28. 325–333. http://proceedings.mlr.press/v28/zemel13.html