What's Individual about Individual Fairness?

Shai Ben-David[1,2], **Pascale Gourdeau**[2,3], **Tosca Lechner**[2,3], **Ruth Urner**[4]

[1]University of Waterloo, Canada
[2]Vector Institute, Canada
[3]University of Toronto, Canada
[4]York University, Canada
shai@uwaterloo.ca, pascale.gourdeau@vectorinstitute.ai, tosca.lechner@vectorinstitute.ai, ruth@eecs.yorku.ca

Abstract

Individual and group fairness notions abound in the machine learning literature. Each attempts to formalize harm against individuals or groups of people. In this work, we take a step back and aim to characterize, from a learning theory perspective, what is at the heart of individual fairness (IF) notions. We argue that fairness notions should be *comparison-based* and, in the case of IF notions, that any failure to be fair should give rise to *finite evidence* of unfairness. We also posit that IF notions should have an unfairness "direction", for example via an order on the set of potential decisions. Equipped with this framework, we present various ways unfair classifiers can be compared to each other. Comparing classifiers is essential in any situation where there is a need to choose between not-perfectly-fair classifiers, e.g., in cases where there exist unavoidable trade-offs between learning objectives. We then adapt score-based measures of individual unfairness to allow us to measure how harm is distributed between population subgroups, which is more in line with group fairness. Crucially, our set-up retains evidence of harm at the individual level, allowing for algorithmic recourse, or potential integrations within legal frameworks.

1 Introduction

As various works have demonstrated the failure of machine learning (ML) models deployed in practice to treat individuals and protected groups fairly (e.g., (Steel and Angwin 2010; Angwin et al. 2016; Obermeyer et al. 2019); see (Mehrabi et al. 2021) for a survey), tremendous efforts have been made by the ML community to mitigate bias and discrimination, both experimentally and theoretically. As a result, the ML community has introduced a profusion of fairness notions to evaluate and train ML models with (Dwork et al. 2012; Hardt, Price, and Srebro 2016; Kearns, Roth, and Wu 2017; Kusner et al. 2017; Kim, Reingold, and Rothblum 2018; Diana et al. 2021). Broadly, these definitions fall into two subcategories: group fairness (GF) and individual fairness (IF). In this work, we focus on individual fairness notions, and faced with this multitude of definitions, we ask,

(RQ1:) *Ultimately, what formally distinguishes individual fairness from other desiderata such as accuracy, explainability and even group fairness?*

Consider a university applicant who gets rejected after applying to their preferred institution. The applicant is understandably upset by the rejection and feels treated "unfairly". Say the student goes on to study at a similar institution and, a couple of years later, completes their degree with distinction. Does this prove that the first university treated this applicant unfairly, or does it merely show that the university made a mistake in assessing the applicant's potential? What type of evidence would solidify the claim of *unfair*, rather simply mistaken, treatment–or are these identical?

Comparison-based evaluation measures. We aim to provide a formal treatment for the above questions in this paper. For this, we start by proposing the formal requirement that any fairness notion should *inherently be based on comparisons between the treatment of different individuals* (1). While the treatment of a single individual can be right or wrong, accurate or inaccurate, this becomes a matter of *fairness* only in relation to the treatment of other individuals. Our technical definition of fairness is related to the philosophical concept of *comparative justice*, which states that "unjust" treatment arises when individuals are treated differently from others.[1] Intuitively, in the above scenario, if the applicant can present a friend from the same school who got accepted despite having lower grades and test scores, their claim of *unfair, and not just wrong*, treatment gets more convincing than if the applicant merely complains about their rejection–even with the later successes at a different institution.

Individual unfairness via sets of finite evidence. Our above-discussed basic requirement of fairness notions being grounded in comparisons (Definition 1) applies to both notions of group fairness and notions of individual fairness. We then provide a property that we argue distinguishes between these two concepts: as opposed to the group notion, individual unfairness should always give rise to *finite evidence* — a bounded-size set of individuals and decisions that demonstrates the violation of the individual fairness notion ("see how *I* got rejected while *they* got accepted with lower grades!"). We provide a formal definition of this distinguishing property in Definition 2 and discuss its relation

[1]See, e.g., (Goodin and Pettit 2019), p.452, for more details, and (Feinberg 1974) for non-comparative justice.

to common notions of fairness from the literature. We want to emphasize that our goal is not to assert which IF notion is best–as this needs to be a case-by-case decision in practice– but rather to distill out the characteristics that let a fairness notion fall into the IF category–what is *individual* about an individual fairness notion? We also postulate that any meaningful IF notion should specify which individuals are being mistreated, for example via an *order* on the set of potential decisions, as the question of *who* is being treated unfairly necessitates a notion of preferential treatment.

Comparing not-perfectly-fair classifiers. While at the individual level, a classifier can be fair or unfair, in practice, because of unavoidable trade-offs between desiderata such as privacy, accuracy, fairness, robustness, etc., one might only be able to choose between *not-perfectly-fair (NPF)* classifiers, begging the question,

(RQ2:) *Given a set of individually NPF classifiers, how should they be evaluated and compared to each other?*

In Section 3, we outline several ways in which NPF classifiers can be compared, via measures/scores of unfairness, or via partial orders. In particular, some of our scores use measures defined at the *individual level*, highlighting the focus on IF and allowing an assessment for each individual.

Structural injustice via evidence sets. Drawing from (Hertweck, Heitz, and Loi 2024), who argued that the way algorithmic fairness and distributive justice have been linked in the literature is too simplistic, we also ask:

(RQ3:) *How can we use our formal framework of evidence sets to study structural injustice?*

In Section 4, we modify scores from Section 3 to measure a classifier's unfairness on population subgroups. Doing so, we effectively extend our IF scores to GF measures, which allows us to compare the (un)fairness of a classifier on different population subgroups to potentially detect structural injustice. We note that, crucially, our approach *maintains evidence of unfairness*, which is in contrast to standard GF notions.

We finally discuss related work in Section 5 as well as limitations in Section 6.

A note on terminology. In this work, we call upon different concepts of justice in political philosophy, e.g., *comparative and non-comparative justice*, *distributive justice*, *structural injustice*, etc. Whenever we mention a specific justice concept, we will define it explicitly. This work ultimately being about fairness from the perspective of learning theory, our use of philosophical concepts is a departure point for our work, and is used to ground our technical contributions in a wider literature. As such, there is no perfect correspondence between our technical contributions and the political philosophy literature, and we do not claim to make novel contributions outside of the field of computer science. We also employ the term *fairness* in a narrower sense than its usage in other computer science and social sciences works; fairness in this work pertains to issues of comparison between individuals in a formal decision-making process. This is addressed in further details in the Ethical Statement section).

2 A Framework for Fairness Notions

2.1 Problem Set-Up

We work in the standard setup of statistical learning: we let $\mathcal{X} \subseteq \mathbb{R}^d$ denote a *feature space*, and let $\mathcal{Y}$ denote the *space of outcomes*. For binary classification, we have $\mathcal{Y} = \{0, 1\}$, for multi-class classification tasks, we may assume $\mathcal{Y} = [k] = \{1, 2, \ldots, k\}$ for a finite number of classes, or $\mathcal{Y} = \mathbb{N}$ for countably many classes. For regression tasks we have $\mathcal{Y} = \mathbb{R}$. In multi-task scenarios the outcome space may be higher dimensional, that is, for each feature vector, we may observe a vector of outcomes for the different tasks.

The data generation is modelled as a distribution P over $\mathcal{X} \times \mathcal{Y}$. When modelling the environment, we refer to such a distribution as the *data-generating distribution* or as the *data-generating process*. We use $P_{\mathcal{X}}$ to denote the marginal of P over $\mathcal{X}$ and P^m for the product distribution corresponding to drawing m i.i.d. samples from P. We denote by $\mathrm{supp}(P)$ the support of a distribution P (and $\mathrm{supp}(P_{\mathcal{X}})$ to refer to the support of the marginal $P_{\mathcal{X}}$). For simplicity, we will usually use statements such as "for all $x \in \mathrm{supp}(P_{\mathcal{X}})$" to express "with probability 1 over $P_{\mathcal{X}}$", and "there exist $x \in \mathrm{supp}(P_{\mathcal{X}})$" to express "with probability greater than 0 over $P_{\mathcal{X}}$". These concepts are equivalent for discrete distributions.

For a binary classification task (i.e., $\mathcal{Y} = \{0, 1\}$), we let $\eta_P : \mathcal{X} \to [0, 1]$ denote the *regression function* of the distribution P, $\eta_P(x) = \Pr_{(x',y)\sim D}[y = 1 | x' = x]$. A *decision rule* is a function $f : \mathcal{X} \to \mathcal{Y}'$, where $\mathcal{Y}'$ is the *decision space*, which may or may not coincide with the space of outcomes $\mathcal{Y}$. For example, in the context of college admissions, we may have $\mathcal{Y} = [0, 1]$, reflecting applicants' scores, while the decision rule (whether or not to admit applicants), must be binary. A *non-deterministic decision rule* is a function $f : \mathcal{X} \to \mathcal{P}(\mathcal{Y}')$, where $\mathcal{P}(\mathcal{Y}')$ denotes the set of probability distributions over $\mathcal{Y}'$. Given a feature space $\mathcal{X}$ and an outcome space $\mathcal{Y}$, we let $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denote the space of distributions over $\mathcal{X} \times \mathcal{Y}$.

Note that by identifying finite datasets with a uniform distribution over their points (taking adequately into account potential repetitions), finite sequences $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ can naturally be viewed as members of $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. We will denote by U_S such distributions on S.

For binary decision rules (i.e., $\mathcal{Y}' = \{0, 1\}$) a non-deterministic decision rule can be identified with a function $f : \mathcal{X} \to [0, 1]$ (with $f(x) \in [0, 1]$ interpreted as a Bernoulli distribution with parameter $f(x)$).

A decision rule is evaluated by means of a *loss function* $\ell : \mathcal{Y}' \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which is point-wise defined on each triple (y', x, y) of decision rule, feature vector and outcome.

For classification tasks, the standard is the binary loss, $\ell^{0/1}(y', x, y) = \mathbb{1}[y' \neq y]$. Typically, the main goal of learning is to identify a decision rule of low *expected loss* $\mathcal{L}_P(f) = \mathbb{E}_{(x,y)\sim P}[\ell(f(x), x, y)]$ over the data-generating distribution P (for non-deterministic f's this expectation is

also over the value of $f(x)$). For the binary loss, this expected loss $\mathcal{L}_P^{0/1}(f)$ coincides with the probability of misclassification.

2.2 Evidence Sets

Aiming to formally distinguish between group-fairness and individual-fairness notions, we start with proposing a formal requirement that any fairness notion should satisfy. First, a notion of fairness assigns an evaluation ("fair" or "unfair") to a decision rule with respect to a data-generating distribution or a finite dataset. Second, we argue that an evaluation of fairness *is based on comparisons between the decisions on different instances*. This crucially differentiates fairness notions from evaluation measures that can be based on a point-wise loss function, such as classification accuracy. To model this distinction, we stipulate that any fairness notion will determine a decision on a single input (for which there are no options for comparisons involving other instances from the feature space) as "fair". The following definition encapsulates this as a requirement. It may be viewed as a *necessary* condition for fairness notions:

Definition 1 (Comparison-based evaluation measure)**.** *Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Y}'$ be a feature, outcome, and decision space respectively, let $\mathcal{F} \subseteq \mathcal{Y}'^{\mathcal{X}}$ be a set of decision rules and let $\mathcal{P} \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be a class of distributions over $\mathcal{X} \times \mathcal{Y}$.*

A comparison-based evaluation measure *for $(\mathcal{X}, \mathcal{Y}, \mathcal{Y}', \mathcal{F}, \mathcal{P})$ is a function $\mathcal{N} : \mathcal{P} \times \mathcal{F} \to \{0, 1\}$ satisfying the following conditions:*

- *If $|\mathrm{supp}(P_{\mathcal{X}})| = 1$ for a distribution $P \in \mathcal{P}$ (that is, only one feature vector is generated by P, and thus no comparisons can be made), then $\mathcal{N}(P, f) = 0$ for all $f \in \mathcal{F}$.*
- *For any distribution $P \in \mathcal{P}$ and decision rules $f, f' \in \mathcal{F}$ that coincide on the support of $P_{\mathcal{X}}$ (that is $\Pr_{x \sim \mathcal{P}_{\mathcal{X}}}[f(x) = f'(x)] = 1$), the evaluation of f and f' with respect to P also coincides, that is $\mathcal{N}(P, f) = \mathcal{N}(P, f')$.*

By default, when not otherwise specified, we let $\mathcal{F} = \mathcal{Y}'^{\mathcal{X}}$ and $\mathcal{P} = \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and then refer to $\mathcal{N}$ as a comparison-based evaluation measure for the triple $(\mathcal{X}, \mathcal{Y}, \mathcal{Y}')$.

We can now propose the above property as a requirement for an evaluation measure that aims at assessing the fairness of a decision rule. We will assume Postulate 1 below in the remainder of this work.

Postulate 1. *A fairness notion is always a comparison based evaluation measure in the sense of Definition 1 above.*

We adopt the convention to interpret $\mathcal{N}(P, f) = 0$ as a verdict of "fair" and $\mathcal{N}(P, f) = 1$ as a determination of "unfair". In Section 2.4 below, we provide examples of how known notions of fairness from the literature fit into the framing of our definition above. In Section 3, we extend the binary notion $\mathcal{N}$ to scores and partial orders in order to be able to compare not-perfectly-fair classifiers, which we apply to structural injustice in Section 4.

We now introduce a requirement that allows us to formally distinguish between group fairness and individual fairness (IF) notions. We argue that a crucial characteristic of individual fairness notions differentiating it from group fairness notions is that it can be attributed to individual instances (in relation to other instances). In particular, when fairness is violated for IF notions, there are specific instances with respect to which an unfair treatment can be claimed, and which then serve as "evidence" of the unfairness. We formalize this principle by introducing a concept of *k-finite evidence of unfairness*.

Definition 2 (Bounded Evidence of Unfairness)**.** *Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Y}'$, $\mathcal{F}$ and $\mathcal{P}$ be as in Definition 1 and let $\mathcal{N}$ be a fairness notion.*

- *We say that $\mathcal{N}$ satisfies the* k-finite evidence property *for some natural number $k \in \mathbb{N}$ if:*

  For every distribution $P \in \mathcal{P}$ and decision rule f with $\mathcal{N}(P, f) = 1$, there exists a finite set K of size at most k in the support of P, such that, for every distribution P' with $K \subseteq \mathrm{supp}(P')$, and every f' that agrees with f on the members of K, we have $\mathcal{N}(P', f') = 1$.
- *We call such a set K an* evidence set *for the unfairness of the decision rule f w.r.t. the data distribution P.*

Throughout the rest of this work, we will denote by $\mathcal{K}_{\mathcal{N},f,P}$ the collection of all evidence sets for $\mathcal{N}(f, P) = 1$ (for fairness notion $\mathcal{N}$, classifier f and distribution P).

Due to the first requirement in Definition 1, the finite evidence K for unfairness will always have size at least 2. This expresses the requirement that fairness is a comparison based evaluation measure. Conceptually, whether a decision on a single individual is correct is a question of non-comparative justice, where decisions made on other individuals are irrelevant.[2]

Note that the above definition is quite general. In particular, it includes cases where fairness of decisions is to be evaluated with respect to the data-generating environment as well as cases where fairness of decisions is to be evaluated with respect to a finite dataset, which is viewed as a uniform distribution over a finite number points labelled with observed outcomes. Note that due to the second requirement in Definition 1, when applied to finite data $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, the outcome of the fairness evaluation of a decision rule f with respect to this data only depends on the labelled points in S and on the values of f on the points x_i occurring in the sequence S. Thus a perhaps more intuitive reading of Definition 2 when applied to finite data is:

> For every sequence $S = ((x_i, y_i, y_i'))_{i=1}^n$ of triples of feature vectors, outcomes and decisions on which fairness is violated, there is a set $T \subseteq S$, of size at most k such that for every set $S' \in (\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}')^m$, for some $m \in \mathbb{N}$, if $T \subseteq S'$ then fairness is also violated on S'.

Moreover, the above finite data-based version of the definition allows us to define a property that we term *outcome-oblivious* for fairness notions. An outcome-oblivious fair-

[2] Again, see, e.g., (Goodin and Pettit 2019), p.452, for more details, and (Feinberg 1974) for non-comparative justice.

ness assessment does not depend on observed outcomes, but rather only on the values the decision rule assigns. Formally:

Definition 3 (Outcome-oblivious fairness notion)**.** *A fairness notion $\mathcal{N}$ is* outcome-oblivious *in case for any $n \in \mathbb{N}$, $T \in \mathcal{X}^n, Y, Y' \in \mathcal{Y}^n$ and any distribution P and decision rule f, $\mathcal{N}(f, U_{(T,Y)}) = \mathcal{N}(f, U_{(T,Y')})$, where (T, Y) and (T, Y') denote the labelled samples with instances T, outcomes Y and decisions Y'.*

2.3 Different vs. Preferential Treatment

Absent from the original definition of individual fairness of Dwork et al. (2012), which deems a decision rule to be t-unfair on $x \in \mathcal{X}$ w.r.t. a metric d if there exists x' with $d(x, x') \leq t$ such that $f(x) \neq f(x')$, is a "direction" to the unfairness: the authors posit that unfairness arises because similar individuals are treated *differently*, rather than an individual receiving a *worse* outcome than another sufficiently similar individual. This view has been transferred to other IF notions inspired by the work of Dwork et al. (2012), e.g., (Lahoti, Gummadi, and Weikum 2019). However, central to legal frameworks as well as algorithmic recourse is the question of *who* is being treated unfairly, or *who* is being harmed by a decision-making process.

Algorithmic recourse has been defined in a variety of ways in the literature, but the overarching idea is that individuals can take actions to change a perhaps unfavourable outcome made by a decision-making model (Ustun, Spangher, and Liu 2019; Joshi et al. 2019; Venkatasubramanian and Alfano 2020). We refer to Karimi et al. (2022) for a survey on the topic. A related concept is that of *algorithmic reparations*, which requires so-called reparative algorithms to "name, unmask, and undo allocative and representational harms as they materialize in sociotechnical form" (Davis, Williams, and Yang 2021). To "name, unmask and undo", one must know which individuals are harmed in a process. Moreover, we note that, in legal cases, plaintiffs must demonstrate with evidence harm done against them. Finally, as we will later see in Section 4, in order to adapt individual fairness scores to structural injustice, we will indeed need a "direction" to unfairness.

To integrate the idea of a preferential treatment to any fairness notion $\mathcal{N}$, we might require that the notion $\mathcal{N}$ identifies a subset of individuals that are mistreated, or harmed, by f for any evidence set K, denoted $K_* \subset K$ (see Postulate 2 below). This subset must be *strict*, as we are comparing individuals, and if someone is worse-off, there must be a better-off individual in the set. For example, in case an IF fairness notion is *outcome-oblivious* (see Definition 3), it suffices to specify an order on the decision space $(\leq, \mathcal{Y}')$, where $y \leq y'$ for $y, y' \in \mathcal{Y}$ signifies that y' is a preferred outcome over y. Then, given an evidence set K and a decision rule f, $(\leq, \mathcal{Y}')$ induces an order on K w.r.t. $f(K)$, where for $x, x' \in K$, $x \leq x'$ iff $f(x) \leq f(x')$, and thus x' gets a better outcome over x. Note that we can also generalize this to a set of orders $\{(\leq_x, \mathcal{Y}')\}_{x \in \mathcal{X}}$, where each individual x has their own preference over outcomes.

For an individual and outcome pair $(x, y) \in K_* \subset K$ for some $K \in \mathcal{K}_{\mathcal{N},f,P}$, the evidence set K (and its mistreated subset K_*) can also be seen as an explanation of why f is unfair on (x, y). We posit that at the core of fairness is the question of mistreatment, and that this should be explicitly incorporated in IF notions, as stated below.

Postulate 2. *A meaningful IF notion satisfying Definitions 1 and 2 should explicitly designate mistreated individuals in all its evidence sets.*

2.4 Common Fairness Notions within our Framework

We will now review a number of group fairness and individual fairness notions and discuss how they are captured by our Definitions 1 and 2 above. This is summarized in Table 1.

Individual Fairness Notions

Similarity-based fairness. This notion, requiring a Lipschitz condition of a decision rule to be fair, was the first fairness notion proposed by Dwork et al. (2012). It is viewed as an IF notion and is adopted by a large body of follow-up work (Ilvento 2020; Dwork, Ilvento, and Jagadeesan 2020; Mahabadi and Vakilian 2020). It formalizes the concept that "similar individuals should be treated similarly", and is based on a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The notion postulates that a decision rule f is considered fair, if for any $x, x' \in \mathcal{X}$: $|f(x) - f(x')| \leq L \cdot d(x, x')$.

The main issue with this notion, already noted by Dwork et al. (2012), is that it requires access to an appropriate notion of similarity, d. In a way, it 'outsources' the problem of defining what fair decisions are to the task of finding a fairness-appropriate similarity function.

If we are considering the fairness of a decision rule f with respect to a distribution P, this boils down to requiring for any $x, x' \in \text{supp}(P_{\mathcal{X}}) \ : \ |f(x) - f(x')| \leq L \cdot d(x, x')$. It is worthwhile noting that the only Lipschitz functions from a continuous connected domain to a discrete range are the constant functions. Therefore, for a data distribution with a connected Euclidean support (e.g., a hypercube or a ball) only constant decision rules meet this fairness requirement. This fairness notion fulfills Definition 2 with $k = 2$, since a violation is demonstrated by just two instances. It does not encompass preferential treatment, as mentioned above, and is also outcome-oblivious.

Lahoti, Gummadi, and Weikum (2019) relaxed the framework of (Dwork et al. 2012) to avoid defining a fairness metric. They proposed a proxy to the measure of (Dwork et al. 2012) by defining a fairness graph, where the nodes are individuals in a training set and the edges encode similarity w.r.t. a given task. Clearly, it also fulfills Definition 2 with sets of size $k = 2$, does not encompass preferential treatment, and is also outcome-oblivious.

Comparison-based fairness. Another IF notion requires that "less qualified" individuals are not preferred over "more qualified" ones. This notion has been introduced as *meritocratic fairness* (Joseph et al. 2016; Kearns, Roth, and Wu 2017; Joseph et al. 2018). Formally, it is based on a (partial) order $\prec$ over the domain $\mathcal{X}$ and a (partial) order $\leq$ over the decision space. A decision rule f is considered unfair if there are individuals x, x' such that $x \prec x'$ and $f(x') \leq f(x)$.

	Fairness notion	Finite evidence?	Outcome-oblivious?	Reflects preferential treatment?
IF	Similarity-based (Dwork et al. 2012)	✓	✓	✗
	Pairwise (Lahoti, Gummadi, and Weikum 2019)	✓	✓	✗
	Meritocratic (Joseph et al. 2016)	✓	✗	✓
	Convexity requirement (this work)	✓	✓	✓
GF	Demographic parity (Dwork et al. 2012)	✗	✓	✗
	Equalized odds (Hardt, Price, and Srebro 2016)	✗	✗	✗
	Equal opportunity (Hardt, Price, and Srebro 2016)	✗	✗	✓

Table 1: Different fairness notions and whether they can be based on finite evidence sets, are outcome-oblivious, and explicitly integrate preferential treatment/direction of unfairness in their definition.

For binary classification with a binary ground truth a natural order to consider here is $x \prec x' \Leftrightarrow \eta_P(x) \leq \eta_P(x')$. Furthermore when judging the fairness of a decision rule with respect to a distribution P (or finite data S), the requirement would be restricted to support elements of $P_{\mathcal{X}}$. With this, the notion also fits into the framework of our Definitions 1 and 2. It satisfies the k-finite evidence property with $k = 2$, since a violation is demonstrated by two instances. As the notion specifies an order on labels, it explicitly integrates preferential treatment in its definition. As natural options for the partial order depend on the ground truth (via the labelling function η_P or observed labels in a finite dataset), we do not consider it outcome oblivious.

Convexity requirement. A convexity requirement can be expressed as follows: the set of instances labelled 1 (or any specified preferred label) by a fair decision rule should be convex. A convexity requirement makes sense when one assumes that the class labels (or the regression function η_ϕ) is monotone in each of the features used in the data representation. In some cases, finding such a representation maybe more likely than finding a representation with respect to which the class labels satisfy a Lipschitz condition. We are not aware of a such fairness notion in the literature, but think it is quite a natural view of IF that could be subject to exploration in the future.

We now show that this requirement satisfies the finite evidence property. If a decision rule h violates this requirement then there are instances $x_1, \dots x_m \in \mathcal{X}$ such that for all $i \leq m$, $h(x_i) = 1$ and some instance $x_{m+1} \in \mathcal{X}$ in the convex hull of $\{x_1, \dots x_m\}$ such that $h(x_{m+1}) = 0$.

While such a violation is evidenced by the set of points $\{x_1, \dots x_m, x_{m+1}\}$, that may have an arbitrary large (finite) size, Carathéodory's theorem states that if $\mathcal{X} \subset \mathbb{R}^d$ then for every set $S \subseteq \mathcal{X}$ and a point $x \in \mathcal{X}$, if x is in the convex hull of S, denoted $\mathrm{conv}(S)$, then there are $x_1, \dots x_{d+1} \in S$ such that x is in the convex hull of these $d + 1$ points. It follows that for $\mathcal{X} \subset \mathbb{R}^d$, violation of this property always has an evidence of size at most $k = d + 2$. Implicitly, the label 1 is here preferred over 0, and the individual x_* labelled 0 in the convex hull can claim unfairness. This notion is outcome-oblivious.

We can also extend the convexity requirement to the multiclass case, $\mathcal{Y} = [l]$ or $\mathcal{Y} = \mathbb{N}$, with the natural order $(\leq, \mathbb{N})$ as follows: for all $S \subset \mathcal{X}$, if $x \in \mathrm{conv}(S)$ then $f(x) \geq \min\{f(x) : x \in S\}$. Intuitively, an individual in the convex hull of S should be treated at least as well as the worse-off individual in S. Thus, a failure is represented as $x \in \mathrm{conv}(S)$ with $f(x) < \min\{f(x) : x \in S\}$, implying as reasoned above there exists $S' \subseteq S$ of size at most $d + 1$ such that $x \in \mathrm{conv}(S')$ with $f(x) < \min\{f(x) : x \in S\} \leq \min\{f(x) : x \in S'\}$.

Group Fairness Notions

Statistical Parity Statistical parity is a classical group fairness notion (Dwork et al. 2012). Based on a partition of the domain into two subgroups A and B, the notion demands that the label proportions of a decision rule in the two subgroups are identical. Thus a decision rule $h : \mathcal{X} \to \{0, 1\}$ is determined to be fair with respect to distribution P over $\mathcal{X} \times \mathcal{Y}$ if $\Pr_{(x,y)\sim P}[h(x) = 1 \mid x \in A] = \Pr_{(x,y)\sim P}[h(x) = 1 \mid x \in B]$. It is easy to see that this notion does not satisfy the k-finite evidence property: for the case of finite datasets, one can always add or a remove a point and change the status between "fair" and "unfair". For GF notions, not integrating preferential treatment is akin to being label invariant, in the sense that applying a permutation to $\mathcal{Y}$ would not change the "fair" or "unfair" judgement. Statistical parity thus does not integrate preferential treatment. It is also outcome oblivious.

Equalized Odds/Equal Opportunity Equalized odds and equal opportunity are two other, by now standard, group fairness notions (Hardt, Price, and Srebro 2016). As statistical parity, these notions hinge on the partition of the domain into two groups A and B. In contrast to statistical parity, they also take observed outcomes (rather than merely decisions) into account. A decision rule $h : \mathcal{X} \to \{0, 1\}$ is determined to be fair in terms of the equal opportunity notion with respect to distribution P over $\mathcal{X} \times \mathcal{Y}$ if $\Pr_{(x,y)\sim P}[h(x) = 1 \mid x \in A, y = 1] = \Pr_{(x,y)\sim P}[h(x) = 1 \mid x \in B, y = 1]$. Equalized odds is a generalization requiring decisions to be independent of group assignments given observed outcomes (Hardt, Price, and Srebro 2016). These notions also do not satisfy the k-finite evidence property. Moreover, as they both depend on the ground truth, they are not outcome oblivious. Equalized odds does not integrate preferential treatment, while equal opportunity, which focuses on positive outcomes, does (there are samples for which we could switch labels and result in a change of "fair" or "unfair" judgment).

3 Comparing Not-Perfectly-Fair Classifiers

In most real life situations, perfectly fair classifiers will be inaccessible. For example, this could be because of some inductive bias of the hypothesis class or noise in the data, or, in the context of *multi-objective* trustworthy machine learning, other criteria such as robustness and explainability may give rise to *unavoidable trade-offs between desiderata*, see, e.g., (Gittens, Yener, and Yung 2022) for a survey. In such situations, how should one *compare* different not-perfectly-fair (NPF) classifiers? In this section, we will propose possible measures of unfairness that could guide a decision on how to choose from a set of NPF classifiers in a way that still attempts to minimize individual harm — even though harm cannot be completely prevented.

We will explore different ways in which evidence sets can be used to compare classifiers from a fairness perspective through (statistical) scores and via partial orders. We outline two philosophies for developing scores: (i) those that arise through a point-wise measure at the level of the *individual* and (ii) those that arise at the level of the the whole *population*. In Section 3.1, we present the individual measures of mistreatment. In Section 3.2, we use these measures to define scores for the classifiers, as well as present the second type of scores, which we call classifier-centred. Finally, in Section 3.3, we outline the partial order perspective.

Recall that, for a set $S \in (\mathcal{X} \times \mathcal{Y})^*$, we denote by U_S the uniform distribution on S (accounting for repetitions), and by $\mathcal{K}_{\mathcal{N},f,P}$ the collection of all evidence sets for $\mathcal{N}(f,P) = 1$ for fairness notion $\mathcal{N}$, classifier f and distribution P.

3.1 Measures of Individual Unfairness

We now start by presenting different options for quantifying the amount of harm an individual suffers from a given decision rule. While we will be using the measures presented here in the following section, we want to emphasize that measuring IF at the individual level is valuable in its own right: it takes the "individual" back into individual fairness, as it allows classifiers to be evaluated and compared from the perspective of the individual. For any two instances (x,y) and (x',y'), the measures presented below allow for deciding whether a decision rule f is better on the former or the latter. Note that all our notions assume the existence of a mistreated set $K_* \subset K$. We now outline these individual-centred measures. The first two can be seen as 0-1 losses, while the other two are probabilities of mistreatment.

- **Individual IF with respect to any distribution:** we define the 0-1 score $\ell_{\mathcal{N}}(x,y,f)$ on a decision rule f and instance (x,y) to be 1 (unfair) in case (x,y) is mistreated in some evidence set K for $\mathcal{N}(f,U_K)=1$:

  $$\ell_{\mathcal{N}}(x,y,f) = \mathbb{1}\left[\exists K \in (\mathcal{X}\times\mathcal{Y})^k : K \in \mathcal{K}_{\mathcal{N},f,U_K} \text{ and } (x,y)\in K_*\right] \ .$$

  Note that for a given distribution P, this might be overly conservative, as evidence sets might not lie in the support for P and thus contain "hypothetical" individuals.
- **Individual IF with respect to distribution P:** we define the 0-1 score $\ell_{\mathcal{N},P}(x,y,f)$ on a decision rule f and instance (x,y) to be 1 (unfair) in case (x,y) is mistreated in some evidence set for $\mathcal{N}(f,P)=1$:

  $$\ell_{\mathcal{N},P}(x,y,f) = \mathbb{1}\left[\exists K \in \mathcal{K}_{\mathcal{N},f,P} \,.\, (x,y)\in K_*\right] \ .$$

  We also note, that similar to before this notion can be adapted to a discrete set of decisions S, by looking at $P = U_S$. It is clear that for any distribution P, we have $\ell_{\mathcal{N},P}(x,y,f) \le \ell_{\mathcal{N}}(x,y,f)$. Thus $\ell_{\mathcal{N}}(x,y,f)=0$ (namely f is fair on (x,y) for any distribution) provides a fairness guarantee for (x,y) for an arbitrary distribution P, namely $\ell_{\mathcal{N},P}(x,y,f)=0$.
- **Probability of drawing evidence of mistreatment:** we define the score $\gamma^K_{\mathcal{N},P}(x,y,f)$ for a given instance (x,y) as the probability of drawing $k-1$ instances to form an evidence set where (x,y) is mistreated:

  $$\gamma^K_{\mathcal{N},P}(x,y,f) \ = \ P^{k-1}(\{K \in \mathcal{K}_{\mathcal{N},f,P} : (x,y)\in K_*\}).$$
- **Probability of having evidence of mistreatment in a sample of size m:** we define the score $\gamma_{\mathcal{N},P,m}(x,y,f)$ for a given instance (x,y) as the probability that there exists an evidence set that contains (x,y) as mistreated when drawing m instances:

  $$\gamma^m_{\mathcal{N},P}(x,y,f) \ = \ \Pr_{S\sim P^m}\left(\exists K' \subset S : K' \cup (x,y) \in \mathcal{K}_{\mathcal{N},f,P} \wedge (x,y)\in K_*\right) \ .$$

3.2 Measures of Population-level Unfairness

We now consider two different types of scores to quantify the fairness of a classifier.

Scores based on individual mistreatment. For any 0-1 individual score $\ell(x,y,f)$ from Section 3.1, it is straightforward to define a population-level measure for a classifier by taking the expectation

$$\Gamma_\ell(f,P) := \mathop{\mathbb{E}}_{(x,y)\sim P}\left[\ell(x,y,f)\right] \ .$$

For probability-based measures $\gamma(x,y,f)$ from Section Section 3.1, we can likewise straightforwardly draw the individual (x,y) from P first. Then, for $\gamma^K_{\mathcal{N},P}$ we define:

$$\Gamma_{\gamma^K_{\mathcal{N},P}}(f,P) := \Pr_{\substack{(x,y)\sim P\\ K\sim P^{k-1}}}\left(K\cup\{(x,y)\}\in\mathcal{K}_{\mathcal{N},f,P} : (x,y)\in K_*\right) \ ,$$

and likewise for the version $\gamma^K_{\mathcal{N},P,m}$ that depends on a sample of size m.

Classifier-centred scores. Above we outlined several ways to define unfairness measures for classifiers based on *aggregating measures of individual mistreatment*. We now also propose options to define measures of unfairness that are not obtained as an expectation of pointwise scores, but rather quantify the overall likelihood of having evidence of unfairness.

- **The probability of evidence sets:** the first score we propose is similar to $\Gamma_{\gamma^K_{\mathcal{N},P}}$, but restated from the classifier point of view, without integrating preferential treatment:

  $$\Gamma^{(1)}_{\mathcal{N}}(f,P) = P^k(\{K\in\mathcal{K}_{\mathcal{N},f,P}\})$$

Note, that the expected number of evidence sets for sample of size m is given by $\binom{m}{k}\Gamma^{(1)}_{\mathcal{N}}(f,P)$.

The second score is analogously defined for a sample of size m:

$$\Gamma^{(2)}_{\mathcal{N}}(f,P,m) := \Pr_{S\sim P^m}(\exists K \subset \mathcal{S} \,.\, K \in \mathcal{K}_{\mathcal{N},f,P}) \quad ,$$

Clearly, $\Gamma^{(1)}_{\mathcal{N}}(f,P) = \Gamma^{(2)}_{\mathcal{N}}(f,P,k)$ and for $m \geq k$, we have that $\Gamma^{(1)}_{\mathcal{N}}(f,P) \leq \Gamma^{(2)}_{\mathcal{N}}(f,P,m)$.

- **Edit-distance to a fair classifier:** After drawing a sample S of size m, this score encapsulates how many labels need to change so that f is fair. For a sample S, let $S|_{\mathcal{X}}$ denote the unlabelled sample and $S|_{\mathcal{Y}}$ the labels. Then we can express this score, denoted $\Gamma^{\text{edit}}_{\mathcal{N}}(f,P,m)$, as

$$\mathop{\mathbb{E}}_{S\sim P^m}\left[\frac{1}{m}\cdot \min_{y\in\mathcal{Y}^m} d_H(y, S|_{\mathcal{Y}}) \mid \mathcal{N}(f,U_S)=0\right] \quad ,$$

where d_H is the Hamming distance.

Interpretation of different types of scores. Broadly speaking, scores based on individual mistreatment can be viewed as aggregating the harm over all individuals in the relevant pool of decisions. Another way to view these scores from the perspective of the decision maker is the *cost of litigation* if individuals had access to the relevant comparisons to argue their case against mistreatment. Every individual that finds evidence for unfairness might be a potential lawsuit. On the other hand, if we look at edit-distance, a classifier-based score, we can interpret this notion as the *cost of changing a decision*. In particular, edit-distance gives an upper bound on the trade-off with potential loss in accuracy, when making a classifier fair.

Individual measures and classifier-centred measures can be misaligned. Consider a similarity-based fairness notion and a set of individuals $S = \{(x_1,y_1),\ldots,(x_m,y_m)\}$ that are pairwise similar. Now consider the the decision rule f with $f(x_1)=1$ and $f(x_2)=\cdots=f(x_m)=0$. We note that while both individual-based and classifier-based scores will attest unfairness in this situation the respective scores given to this decision rule are very different:

- Consider $\ell_{\mathcal{N},U_S}(f,x,y)$. We can see that for every $i \in \{2,\ldots,m\}$ we have $\ell_{\mathcal{N},U_S}(f,x,y) = 1$, as $K_i = \{(x_1,y_1),(x_i,y_i)\} \in \mathcal{K}_{\mathcal{N},U_S,f}$ and $(x_i,y_i) \in K_i^*$. Thus, $\Gamma_{\ell_{\mathcal{N},U_S}} = \frac{m-1}{m}$.
- On the other hand, consider edit distance. It is easy to see that the classifier f' with $f'(x_1)=\cdots=f'(x_m)=0$ is fair and has Hamming-distance 1 to f. Thus for every m, we have $\Gamma^{\text{edit}}_{\mathcal{N}}(f,U_S,m) \leq \frac{1}{m}$

The above example shows that it is sometimes easy to prevent a considerable number of potential legal actions by adapting the decision rule on only a few examples. Crucially however, this adaptation is done in a way that gives a less favourable outcome to one individual, while not changing the treatment for any other individual. Thus, from a perspective of welfare, the original decision rule f was Pareto-dominating the fairness-adapted decision rule f'.

3.3 Comparing Unfairness through Evidence Set Inclusion

Let $\mathcal{F}$ be the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$. We define a partial order $(\mathcal{F}, \leq_{\mathcal{N}})$ where $f \leq_{\mathcal{N}} g$ iff $\mathcal{K}_{\mathcal{N},f,P} \subseteq \mathcal{K}_{\mathcal{N},g,P}$. Namely $f \leq_{\mathcal{N}} g$ in case any evidence set for f is also an evidence set for g. Then, if $f \leq_{\mathcal{N}} g$, we can argue that f is preferable to g as follows: on the population level, if $f \leq_{\mathcal{N}} g$ then for both types of population level scores the following inequalities hold: $\Gamma^{(1)}_{\mathcal{N}}(f,P) \leq \Gamma^{(1)}_{\mathcal{N}}(g,P)$ and $\Gamma^{(2)}_{\mathcal{N}}(f,P,m) \leq \Gamma^{(2)}_{\mathcal{N}}(g,P,m)$ for all sample sizes m (but the other direction does not necessarily hold). Thus from a statistical (population level) perspective, f is preferable to g. From the individuals' perspectives, notate that if Postulate 2 is satisfied, $f \leq_{\mathcal{N}} g$ means that the set of individuals mistreated by f is contained in the set of individuals mistreated by g, i.e. $\{x \in \mathcal{X} \mid \exists K \in \mathcal{K}_{\mathcal{N},f,P} \,.\, x \in K_*\} \subseteq \{x \in \mathcal{X} \mid \exists K \in \mathcal{K}_{\mathcal{N},g,P} \,.\, x \in K_*\}$. Thus we established a preference of f over g at the individual level as well, retaining the *individuality* component of IF.

Clearly, this approach has the drawback that some classifiers are incomparable because neither $\mathcal{K}_{\mathcal{N},f,P} \subseteq \mathcal{K}_{\mathcal{N},g,P}$ nor $\mathcal{K}_{\mathcal{N},g,P} \subseteq \mathcal{K}_{\mathcal{N},f,P}$ holds. We note however that the above outlined approach of comparing classifiers through inclusion of evidence sets could play an important role for explainability in multi-objective trustworthy ML: given classifiers f,g, the partial order defined through evidence set inclusion offers an explanation of why f was chosen over g.

4 Structural Injustice via Evidence Sets

Fairness is often invoked in scenarios where we want to know whether membership to a certain group increases the chances of mistreatment. Group fairness (GF) notions (see Section 2.4 for examples) are directly aimed at measuring this. In this section, we will discuss how individual fairness (IF) and our evidence set based scores to measure mistreatment from Section 3 above can be used to also derive scores that could detect structural injustice. Additionally, later in this section, we will highlight a conceptual link between the work of Hertweck, Heitz, and Loi (2024) and our approach.

As argued in Section 2.3, from an algorithmic recourse and reparation perspective, identifying individuals that are unfairly treated is important. We now argue that basing scores for structural injustice on evidence of individual mistreatment as we here suggest allows for simultaneously flagging unfairness at the group level and pointing out *specific individuals* who have been harmed by an unfair treatment from the classifier – something the traditional group fairness measures presented in Section 2.4 cannot do. Indeed, a "proof" of unfairness at the individual level for scores like equalized odds or equal opportunity (Hardt, Price, and Srebro 2016) could only consist of a misclassification on a single instance. This does not satisfy our criterion of Definition 2 that evidence sets must have size at least 2 and that fairness is inherently a comparison based property.[3]

[3]This individual criterion would be more in line with non-comparative justice, where other individuals are irrelevant. Again, see the work of Parfit (2018) for details. We note that Kamishima

This reflects a common struggle of members of marginalized groups: even when it is acknowledged that an individual belongs to a disadvantaged group, how does the individual argue that a specific treatment has been *unfair*? Common group fairness notions do not provide any avenue for an individual to make such an argument. Grounding measures of structural injustice in evidence based scores of individual mistreatment on the other hand, would provide a pathway to recourse for the members of disadvantaged groups.

Scores for structural injustice. The definitions and constructions from Section 3 allow for classifiers to be compared with fairness scores, in the sense that we can quantify how much a classifier violates individual fairness. We can extend these fairness scores to to reflect group-based discrimination, given our set-up of mistreated sets $K_* \subset K$ for an evidence set K.

Let $A = \{1, \ldots, l\}$ be an attribute (race, gender, marital status, etc.), and given a classifier f, a distribution P and an index $a \in A$ of group membership. The first approach we presented in Section 3.1, which focused on *individual measures*, allows us to do this easily by taking the expectation over the conditional distribution P_a for each $a \in A$. Namely, all population-level measures based on individual mistreatment in Section 3.2 can be evaluated with respect to the conditional distributions P_a, and these scores can be compared between groups.

Note that the existence of an order $(\leq, \mathcal{Y})$ or a mistreated set $K_* \subset K$ is crucial to define structural injustice as demonstrated below.

Proposition 1. *Any fairness score arising from a fairness notion $\mathcal{N}$ used to detect structural injustice must satisfy Postulate 2.*

Proof. Let $(x_1, y_1), (x_2, y_2) \in \mathcal{X} \times \mathcal{Y}$, and let distribution P be such that the probability of drawing (x_1, y_1) is $p > 0$ and is equal to that of drawing (x_2, y_2). Let $\mathcal{N}^{\text{no-pref}}$ be an individual fairness notion satisfying Definitions 1 and 2, but not Postulate 2 (it does not designate mistreated individuals). Let f be such that $\mathcal{N}(f, P) = 1$ and there is only one evidence set: $\mathcal{K}_{\mathcal{N},f,P} = \{K\}$, where $K = \{(x_1, y_1), (x_2, y_2)\}$. Let $A = \{0, 1\}$ represent two groups, and suppose x_1 belongs to 0 and x_2 to 1, and moreover that P_0 and P_1 assign the same probability p' to x_1 and x_2, respectively. Then, by symmetry, any group measure of unfairness $\mathcal{N}^{\text{no-pref}}$ will be the same for $a = 0$ or 1. However, satisfying Postulate 2 implies that the two measures are different, e.g., for $\Gamma_{\gamma^K_{\mathcal{N},P}}$ defined in Section 3.2, if $(x_1, y_1) = K_*$, we have that $\Gamma_{\gamma^K_{\mathcal{N},P}}(f, P, 0) = p'p$ and $\Gamma_{\gamma^K_{\mathcal{N},P}}(f, P, 1) = 0$. □

Relationship with (Hertweck, Heitz, and Loi 2024). The work of Hertweck, Heitz, and Loi (2024) studies how algorithmic fairness and theories of distributive justice have been linked in the literature (often too simplistically). They ask,

> If theories of distributive justice are defined at the level of individuals, how do they relate to group fairness criteria defined at the level of socio-demographic groups?

In their work, Hertweck, Heitz, and Loi (2024) refer to *distributions* as the way decisions are made/how harm or benefits are allocated; *ideals* refer to how these decisions ought to be made under some moral principles (which can be fulfilled by multiple *ideal distributions*). A *deviation from the ideal* is the set of decisions that contradict the ideal. We will now outline some conceptual connections between these and our work. In our framework, the ideal can be seen as the principle behind the fairness notion $\mathcal{N}$ (ideal distributions being the pairs f, P under which $\mathcal{N}(f, P) = 0$), and the deviation from the ideal is the set $\mathcal{K}_{\mathcal{N},f,P}$ of evidence sets causing $\mathcal{N}(f, P) = 1$. In our view, theories of distributive justice are naturally related to IF notions, which evaluate the quality of decisions made on individuals. Structural injustice, which is what GF notions aim to evaluate, is defined by Hertweck, Heitz, and Loi (2024) as how there can be bias in how "the deviations from the ideal... affect some groups more than others". We acknowledge that our interpretation of the work of Hertweck, Heitz, and Loi (2024) is subjective. We have used it as a departure point to study structural injustice within our framework of evidence sets. As such, there is no perfect correspondence between our problem set-up and definitions and the theory presented by Hertweck, Heitz, and Loi (2024).[4] We view the latter as informing and situating our work.

5 Related Work

The notion of individual fairness was first introduced in (Dwork et al. 2012), motivated by the idea that *similar individuals should be treated similarly*. This idea was then formalized as a requirement on the Lipschitzness of the decision rule, given a metric that captures *task relevant similarity*. The notion of similarity-based individual fairness has since been extended into different domains of machine learning, such as clustering (Mahabadi and Vakilian 2020), graph mining (Kang et al. 2020), graph neural networks (Dong et al. 2021). One issue with this notion is that such a task relevant metric is in general not known to the learner — if it was, learning itself would not be as hard a task. In contrast, when similarity-based individual fairness is applied, the task-relevant metric often is just assumed to be known, with practitioners just using an underlying easily available metric, without arguing for it capturing task relevance. In many cases the output layer is used as a proxy for the task-relevant metric (Dong et al. 2021), which often leads to only post-hoc justifying the learned labelling rule with a learned precursor to the decision rule, rather then investigating its fairness independently.

One attempt to overcome this issue has been to *learn* the fairness metric from expert advice (Ilvento 2020). However such attempts often rely on imprecise feedback as similarity

(2023) extracted individual criteria for group fairness measures, which are more in line with non-comparative justice criteria.

[4] See, e.g., their distinction between distributive justice criterion and metric, and fairness criterion and metric. Moreover, not every IF notion within our framework can necessarily be traced back to a theory of distributive justice.

is hard to subjectively quantify, as well as on sufficient prior knowledge of a class of relevant fairness metrics.

Another issue that has been addressed is quantifying unfairness and learning an individually fair classifiers. An approximate and probabilistic notion of metric-fairness, similar to the notion of probabilistic Lipschitzness (Urner, Wulff, and Ben-David 2013), is used in (Rothblum and Yona 2018) to define a version of metric fair probably approximately correct (PAC) learning. Similarly, the more general notion of metric multi-fairness (Kim, Reingold, and Rothblum 2018) addresses this problem by restricting the collection of sets from which comparisons are drawn and then giving a worst-case guarantee amongst average comparisons *within* each set. In contrast to the notions we propose, which are based on finite evidences, these notions quantify fairness probabilistically with respect to either a distribution or a set of uniform distributions given a collection of sets. In order for unfairness to be inferred, the structure of the hypothesis class or collection of comparison sets needs to be restricted, and guarantees can only be given statistically. In comparison, our notion is more directly tied to the actual decisions being made for a group of individuals, thus more directly quantifying fairness for individuals.

There have also been so-called "metric-free" approaches to individual fairness. Lahoti, Gummadi, and Weikum (2019) define a fairness graph with individuals as nodes and edges representing pairs of individuals deemed to be sufficiently similar by a human expert or judge, and use this graph to learn pairwise fair representations. In a similar vein, instead of having access to the task-relevant metric, Gillen et al. (2018); Bechavod, Jung, and Wu (2020); Bechavod (2024) assume that there is access to a fairness judgment of decisions in an online learning setting. The fairness notion here is still assumed to come from a metric-based notion, but the information of violations is only determined by the judge, who judges on a finite number of made decisions. In this regard the last two notions bare resemblance to our finite-evidence criterion for individual fairness notions. The online-learning protocol proposed in (Bechavod, Jung, and Wu 2022) can likely be generalized to all fairness notions fulfilling our finite-evidence set criterion. In particular, a generalization has been shown that for general abstract individual fairness notions that are based on the comparisons of two individuals (Jung et al. 2021), allowing also for order-based comparisons. In another setting of sequential decision making (Gupta and Kamble 2021) propose a distinction between *fairness over time*, which proposes that the set of all decisions made in the sequence satisfy the similarity-based individual fairness notion, and *fairness in hindsight*, which only requires decisions to be compared to decisions that have already been made. These notions can also be generalized to other fairness notions fulfilling our finite-evidence property. A different notion in the literature that was also termed "metric-free individual fairness" is one that postulates pointwise independence of decisions and sensitive attributes.

Over the years there have been a number of criticisms of metric-based individual fairness. Fleisher (2021) discusses the philosophical insufficiencies of metric-based fairness as well as its inapplicability due to lack of access to the metric. Furthermore, it has been found that there is a trade-off between group and individual fairness (Xu and Strohmer 2024) and that metric-based individual fairness can be gerrymandered (Räz 2024). It has also been proposed that for many tasks, partial orders might be a more natural structure with which to compare two individuals (Jung et al. 2021; Fleisher 2021) to judge the fairness of a decision, which also fits in our evidence-set treatment of individual fairness. Beyond the critiques of individual fairness, different fairness frameworks have been explored in relation to each other and world beliefs (Friedler et al. 2019; Friedler, Scheidegger, and Venkatasubramanian 2021) and in relation to theories of distributive justice (Binns 2018; Kuppler et al. 2021; Hertweck, Heitz, and Loi 2024) and more precisely egalitarianism (Mittelstadt, Wachter, and Russell 2023). Other taxonomies and formalizations of fairness notions have been proposed in the literature. For example, Kamishima (2023) re-formalizes individual and group fairness criteria, taking statistical GF measures and extracting an IF criterion for each, and Corbett-Davies et al. (2023) separate fairness definitions into two categories, depending on whether they limit the effects of decisions on disparities or of protected attributes on decisions.

Finally, fairness measures have been found to suffer from limitations and incompatibilities (Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2017; Lechner and Ben-David 2022; Corbett-Davies et al. 2023), and there has been substantial interest in developing methods to satisfy fairness requirements (Donini et al. 2018; Madras et al. 2019; Liu, Simchowitz, and Hardt 2019; Dwork, Ilvento, and Jagadeesan 2020; Slack, Friedler, and Givental 2020; Xian, Yin, and Zhao 2023) and verifying fairness (Ruoss et al. 2020; Benussi et al. 2022; Wicker, Piratla, and Weller 2023; Yadav et al. 2024).

6 Limitations

With respect to our framework, evidence sets and fairness notions do not give any indication on their own about how well a fairness construct fits a specific problem. While they formally distill out necessary requirements on fairness notions, they certainly do not provide sufficient ones. Scores as we defined in Section 3 can also be seen as departing from core tenets of individual fairness proposed in Dwork et al. (2012).

7 Discussion & Conclusion

Our work introduces a new perspective on individual fairness, which we hope might help focus and advance future discussion on the concept. In particular, we postulated being grounded in comparisons as necessary requirement for all fairness notions, and we identified a property, namely being based on finite evidence, that distinguishes between group and individual notions of unfairness. We also argued for a broader view of individual fairness beyond Lipschitz-requirements. We proposed ways to employ the notion of finite evidence for unfairness to define scores and partial orders, allowing comparing classifiers that are not perfectly

fair. We developed scores first through measures of unfairness at the individual level, which are valuable in their own right. We adapted these scores to reflect structural injustice, drawing on interdisciplinary work in algorithmic fairness and philosophy. We believe that, while having made technical points, the conceptual part of our work and its positioning in the broader fairness and justice literature is its most significant contribution.

Ethical Statement

Before integrating fairness or other trustworthy objectives in a ML system, one must first make a convincing argument that using ML for a given problem is ethical or desirable in the first place (Wang et al. 2024); addressing harm within the ML system via evaluation metrics may not achieve any sort of justice, and using any automated decision making system may even worsen inequities. As argued by many in the algorithmic fairness literature, algorithmic fairness can fail to capture intricacies of real-world scenarios and is thus more times than not simply a way to raise a red flag on potential discrimination, harm or unfairness — passing some algorithmic fairness test does not necessarily imply fairness in a broader sense, and justice even less. It is even possible for entities to "fairwash" their model by hiding unfair behaviour while still passing a fairness audit (Shahin Shamsabadi et al. 2022). Moreover, our definitions do not address systemic issues and power structures at the root of injustice, as is the case with most if not all mathematical framings of algorithmic fairness, which only locally tackle its effects (Kasirzadeh 2022). Kasirzadeh (2022) argues that the problem of structural injustice

> ...requires widening the scope of algorithmic fairness to include power relations, social dynamics and actors and structures which are among the main sources of the emergence and persistence of social injustices relevant to algorithmic systems.

In the same line of thought, the idea "ground truth", usually a given in learning theory, can be intangible in practice, especially in the context of fairness. Moreover, the observable labels and data distribution could be biased in a way that is not addressable within our framework (the work of Fazelpour and Lipton (2020) studies this from the perspective of philosophy perspectives, drawing on "ideal" and "non-ideal" methodological approaches). It has later been attempted by Dwork, Reingold, and Rothblum (2023) to distinguish between the "real and ideal worlds" in a mathematical way. However, the positive results therein can be guaranteed if and only if robustness assumptions on the set of potential transformations from the real to the ideal world satisfy a type of robustness, which does not necessarily hold in practice.

Acknowledgements

Shai Ben-David is a Vector faculty member and a CIFAR AI chair and he thanks both Vector and CIFAR for their support of this research. Pascale Gourdeau has been supported by a Vector Postdoctoral Fellowship and an NSERC Postdoctoral Fellowship. Tosca Lechner has been supported by a Vector Postdoctoral Fellowship. Ruth Urner is also an Affiliate Faculty Member at Toronto's Vector Institute, and her research is supported by an NSERC discovery grant.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias.

Bechavod, Y. 2024. Monotone individual fairness. In *Proceedings of the 41st International Conference on Machine Learning*, 3266–3283.

Bechavod, Y.; Jung, C.; and Wu, Z. S. 2020. Metric-Free Individual Fairness in Online Learning. *Advances in neural information processing systems*, 33: 11214–11225.

Bechavod, Y.; Jung, C.; and Wu, Z. S. 2022. Metric-Free Individual Fairness in Online Learning. arXiv:2002.05474.

Benussi, E.; Patane, A.; Wicker, M.; Laurenti, L.; and Kwiatkowska, M. 2022. Individual Fairness Guarantees for Neural Networks. In *31st International Joint Conference on Artificial Intelligence, IJCAI 2022*, 651–658. International Joint Conferences on Artificial Intelligence (IJCAI).

Binns, R. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, 149–159. PMLR.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2): 153–163.

Corbett-Davies, S.; Gaebler, J. D.; Nilforoshan, H.; Shroff, R.; and Goel, S. 2023. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312): 1–117.

Davis, J. L.; Williams, A.; and Yang, M. W. 2021. Algorithmic reparation. *Big Data & Society*, 8(2): 20539517211044808.

Diana, E.; Gill, W.; Kearns, M.; Kenthapadi, K.; and Roth, A. 2021. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 66–76.

Dong, Y.; Kang, J.; Tong, H.; and Li, J. 2021. Individual Fairness for Graph Neural Networks: A Ranking based Approach. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, 300–310. Association for Computing Machinery. ISBN 9781450383325.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; and Pontil, M. 2018. Empirical Risk Minimization Under Fairness Constraints. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2796–2806.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science conference*, 214–226.

Dwork, C.; Ilvento, C.; and Jagadeesan, M. 2020. Individual Fairness in Pipelines. In Roth, A., ed., *1st Symposium on*

Foundations of Responsible Computing, FORC, volume 156 of *LIPIcs*, 7:1–7:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Dwork, C.; Reingold, O.; and Rothblum, G. N. 2023. From the real towards the ideal: Risk prediction in a better world. In *4th Symposium on Foundations of Responsible Computing (FORC 2023)*, 1–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Fazelpour, S.; and Lipton, Z. C. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 57–63.

Feinberg, J. 1974. Noncomparative Justice. In *Philosophical Review*, volume 83.

Fleisher, W. 2021. What's fair about individual fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 480–490.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4): 136–143.

Friedler, S. A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E. P.; and Roth, D. 2019. A comparative study of fairness-enhancing interventions in machine learning. In danah boyd; and Morgenstern, J. H., eds., *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT**, 329–338. ACM.

Gillen, S.; Jung, C.; Kearns, M.; and Roth, A. 2018. Online learning with an unknown fairness metric. *Advances in neural information processing systems*, 31.

Gittens, A.; Yener, B.; and Yung, M. 2022. An adversarial perspective on accuracy, robustness, fairness, and privacy: multilateral-tradeoffs in trustworthy ML. *IEEE Access*, 10: 120850–120865.

Goodin, R. E.; and Pettit, P. 2019. *Contemporary political philosophy: an anthology*. John Wiley & Sons.

Gupta, S.; and Kamble, V. 2021. Individual Fairness in Hindsight. *Journal of Machine Learning Research*, 22(144): 1–35.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems NIPS*, 3315–3323.

Hertweck, C.; Heitz, C.; and Loi, M. 2024. What's Distributive Justice Got to Do with It? Rethinking Algorithmic Fairness from a Perspective of Approximate Justice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 597–608.

Ilvento, C. 2020. Metric Learning for Individual Fairness. In Roth, A., ed., *1st Symposium on Foundations of Responsible Computing, FORC*, volume 156 of *LIPIcs*, 2:1–2:11. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Joseph, M.; Kearns, M. J.; Morgenstern, J.; Neel, S.; and Roth, A. 2018. Meritocratic Fairness for Infinite and Contextual Bandits. In Furman, J.; Marchant, G. E.; Price, H.; and Rossi, F., eds., *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, 158–163. ACM.

Joseph, M.; Kearns, M. J.; Morgenstern, J.; and Roth, A. 2016. Fairness in Learning: Classic and Contextual Bandits. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, NIPS*, 325–333.

Joshi, S.; Koyejo, O.; Vijitbenjaronk, W.; Kim, B.; and Ghosh, J. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*.

Jung, C.; Kearns, M.; Neel, S.; Roth, A.; Stapleton, L.; and Wu, Z. S. 2021. An Algorithmic Framework for Fairness Elicitation. In *2nd Symposium on Foundations of Responsible Computing*, volume 31, 21.

Kamishima, T. 2023. Re-formalization of Individual Fairness. *arXiv preprint arXiv:2309.05521*.

Kang, J.; He, J.; Maciejewski, R.; and Tong, H. 2020. InFoRM: Individual Fairness on Graph Mining. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20.

Karimi, A.-H.; Barthe, G.; Schölkopf, B.; and Valera, I. 2022. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5): 1–29.

Kasirzadeh, A. 2022. Algorithmic fairness and structural injustice: Insights from feminist political philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 349–356.

Kearns, M. J.; Roth, A.; and Wu, Z. S. 2017. Meritocratic Fairness for Cross-Population Selection. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML*, volume 70 of *Proceedings of Machine Learning Research*, 1828–1836. PMLR.

Kim, M.; Reingold, O.; and Rothblum, G. 2018. Fairness through computationally-bounded awareness. *Advances in neural information processing systems*, 31.

Kleinberg, J. M.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. H., ed., *8th Innovations in Theoretical Computer Science Conference, ITCS*, volume 67 of *LIPIcs*, 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Kuppler, M.; Kern, C.; Bach, R. L.; and Kreuter, F. 2021. Distributive justice and fairness metrics in automated decision-making: How much overlap is there? *arXiv preprint arXiv:2105.01441*.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Lahoti, P.; Gummadi, K. P.; and Weikum, G. 2019. Operationalizing individual fairness with pairwise fair representations. *Proceedings of the VLDB Endowment*, 13(4): 506–518.

Lechner, T.; and Ben-David, S. 2022. Inherent Limitations of Multi-Task Fair Representations. In Chandar, S.; Pascanu, R.; and Precup, D., eds., *Conference on Lifelong Learning Agents CoLLAs*, volume 199 of *Proceedings of Machine Learning Research*, 583–603. PMLR.

Liu, L. T.; Simchowitz, M.; and Hardt, M. 2019. The Implicit Fairness Criterion of Unconstrained Learning. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97, 4051–4060. PMLR.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. S. 2019. Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data. In danah boyd; and Morgenstern, J. H., eds., *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, 349–358. ACM.

Mahabadi, S.; and Vakilian, A. 2020. Individual Fairness for k-Clustering. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6586–6596. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Mittelstadt, B.; Wachter, S.; and Russell, C. 2023. The Unfairness of Fair Machine Learning: Leveling Down and Strict Egalitarianism by Default. *Mich. Tech. L. Rev.*, 30: 1.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Parfit, D. 2018. Equality and priority 1. In *The Notion of Equality*, 427–446. Routledge.

Rothblum, G. N.; and Yona, G. 2018. Probably Approximately Metric-Fair Learning. arXiv:1803.03242.

Ruoss, A.; Balunovic, M.; Fischer, M.; and Vechev, M. 2020. Learning certified individually fair representations. *Advances in neural information processing systems*, 33: 7584–7596.

Räz, T. 2024. Gerrymandering individual fairness. *Artificial Intelligence*, 326: 104035.

Shahin Shamsabadi, A.; Yaghini, M.; Dullerud, N.; Wyllie, S.; Aïvodji, U.; Alaagib, A.; Gambs, S.; and Papernot, N. 2022. Washing the unwashable: On the (im) possibility of fairwashing detection. *Advances in Neural Information Processing Systems*, 35: 14170–14182.

Slack, D.; Friedler, S. A.; and Givental, E. 2020. Fairness warnings and fair-MAML: learning fairly with minimal data. In Hildebrandt, M.; Castillo, C.; Celis, L. E.; Ruggieri, S.; Taylor, L.; and Zanfir-Fortuna, G., eds., *FAT* '20: Conference on Fairness, Accountability, and Transparency*, 200–209. ACM.

Steel, E.; and Angwin, J. 2010. On the web's cutting edge, anonymity in name only. *The Wall Street Journal*, 4.

Urner, R.; Wulff, S.; and Ben-David, S. 2013. PLAL: Cluster-based active learning. In Shalev-Shwartz, S.; and Steinwart, I., eds., *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, 376–397. Princeton, NJ, USA: PMLR.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, 10–19.

Venkatasubramanian, S.; and Alfano, M. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 284–293.

Wang, A.; Kapoor, S.; Barocas, S.; and Narayanan, A. 2024. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Journal on Responsible Computing*, 1(1): 1–45.

Wicker, M.; Piratla, V.; and Weller, A. 2023. Certification of distributional individual fairness. *Advances in Neural Information Processing Systems*, 36: 27670–27681.

Xian, R.; Yin, L.; and Zhao, H. 2023. Fair and Optimal Classification via Post-Processing. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML*, volume 202 of *Proceedings of Machine Learning Research*, 37977–38012. PMLR.

Xu, S.; and Strohmer, T. 2024. On the (In)Compatibility between Group Fairness and Individual Fairness. arXiv:2401.07174.

Yadav, C.; Chowdhury, A. R.; Boneh, D.; and Chaudhuri, K. 2024. FairProof: Confidential and Certifiable Fairness for Neural Networks. In *International Conference on Machine Learning*, 55682–55705. PMLR.