A Unified View of Group Fairness Tradeoffs Using Partial Information Decomposition

Faisal Hamman and Sanghamitra Dutta
Department of Electrical and Computer Engineering, University of Maryland College Park
Emails: fhamman@umd.edu, sanghamd@umd.edu

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619698

***Abstract*—This paper introduces a novel information-theoretic perspective on the relationship between prominent group fairness notions in machine learning, namely statistical parity, equalized odds, and predictive parity. It is well known that simultaneous satisfiability of these three fairness notions is usually impossible, motivating practitioners to resort to approximate fairness solutions rather than stringent satisfiability of these definitions. However, a comprehensive analysis of their interrelations, particularly when they are not exactly satisfied, remains largely unexplored. Our main contribution lies in elucidating an exact relationship between these three measures of (un)fairness by leveraging a body of work in information theory called partial information decomposition (PID). In this work, we leverage PID to identify the granular regions where these three measures of (un)fairness overlap and where they disagree with each other leading to potential tradeoffs. We also include numerical simulations to complement our results.**

I. Introduction

The rapid infiltration of machine learning (ML) into high-stakes applications such as employment, education, finance, healthcare brings the promise of enhanced efficiency. However, this can also accompanied by escalating concerns about the disparate impact [1]–[5] that these systems might cause on unprivileged *groups* based on sensitive attributes such as gender, race, age, nationality, etc. Several anti-discrimination legislations and ethical principles [1] are being actively put forth to ensure algorithmic fairness.

Existing literature has fostered a plethora of definitions, metrics, and scholarly debates about algorithmic fairness [6]. Central to the debate of quantifying fairness at a group level are three popular definitions, namely, statistical parity, equalized odds, and predictive parity [6]–[8]. Due to the multitude of fairness definitions available, it is often unclear which measure of fairness is most appropriate to adopt in a given setting [9]. Furthermore, it is also well-known that simultaneous satisfiability of these three fairness definitions is generally impossible [10], [11].

Given such a fundamental impossibility, practitioners often strive for approximate fairness solutions rather than stringent satisfiability of all these definitions. Such approximate fairness solutions consist of two pivotal aspects: (i) quantification of (un)fairness (i.e., a gap from exact satisfiability); and (ii) development of strategies to mitigate such unfairness in ML models. For instance, one may jointly minimize one or more measures of unfairness while training an ML model which has often led to empirical tradeoffs between accuracy and different measures of unfairness [12], [13].

Although previous studies have identified certain impossibilities among these fairness notions [10], [11], a detailed analysis focusing on *the interrelationships among different measures of unfairness*, specifically explaining when they will be in agreement and when they will be in disagreement leading to potential tradeoffs has received limited attention.

Our research bridges this gap by leveraging Partial Information Decomposition (PID) [14], a body of work in information theory, to elucidate the exact relationship between different measures of unfairness. In particular, we consider information-theoretic quantifications [15] of the respective gaps from statistical parity, equalized odds, and predictive parity as our measures of unfairness. Using PID, we demonstrate the exact relationship between these three measures of unfairness in Proposition 1 (also see Fig. 3 for a pictorial illustration of the relationship between the measures of unfairness).

PID enables us to provide a unified information-theoretic framework that is instrumental in establishing the fundamental limits and tradeoffs among these unfairness measures, particularly in the context of approximate fairness solutions when exact satisfiability of all three fairness definitions is not met. Furthermore, the impossibility among the three fairness definitions can also be derived from our result (see Theorem 1). We also identify and delineate the regions of agreement and disagreement among these three measures of unfairness (see Section III), providing insights on when there will be a tradeoff and when there will be no tradeoff among the measures of unfairness. We perform numerical simulations on the Adult dataset [16] to complement our theoretical results. Moreover, our work holds broader implications in fields such as algorithmic fairness auditing [17], where it can significantly contribute to the evaluation of fairness in ML models.

Related Works: Information-theoretic measures have been used to study group fairness in the fairness literature [4], [5], [15], [18]–[27]. Another related line of work is exploring trade-offs between fairness and accuracy [12], [28]–[32].

PID is recently gaining traction across various ML applications [4], [5], [33]–[40]. It is particularly noteworthy in the realm of algorithmic fairness [4], [5], [33], [34]. [5] leverages PID to dissect total disparity in decision-making into exempt and non-exempt components. Similarly, [34] employs PID to study the interplay between global and local fairness in federated learning. We also refer to [33] for a survey of PID in fairness and explainability. Understanding tradeoffs and agreement disagreement between unfairness measures using

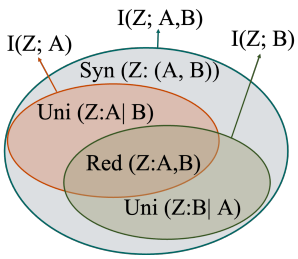

Fig. 1. Venn diagram showing PID of $\mathrm{I}(Z; A, B)$.

PID has not been studied. In this work, our objective is to develop a unified information-theoretic framework that effectively delineates the fundamental limits and trade-offs among the existing unfairness measures: statistical parity, equalized odds, and predictive parity.

II. Preliminaries

Let X denote the input features, Z denote the sensitive attribute, and Y denote the true label. The sensitive attribute Z is assumed to be binary with 1 indicating the privileged group and 0 indicating the unprivileged group. We also let $\hat{Y}$ represent the predictions of a model, i.e., $\hat{Y} = f_\theta(X)$ where the model is parameterized by θ. Standard machine learning aims to minimize the empirical risk:

$$\min_\theta L(\theta) = \min_\theta \frac{1}{n} \sum_{i=1}^{n} l(f_\theta(x_i), y_i),$$

where $l(\cdot,\cdot)$ is a predefined loss function, x_i is the input feature, $y_i \in \{0,1\}$ is the true label, and n is the number of datapoints in the dataset.

A. Background on Partial Information Decomposition

Partial Information Decomposition (PID) [14] decomposes the total mutual information about a random variable Z contained in the tuple (A, B), i.e., $\mathrm{I}(Z; A, B)$ into four *nonnegative* terms as follows (also see Fig. 1):

$$\begin{aligned}\mathrm{I}(Z; A, B) = \mathrm{Uni}(Z{:}A|B) + \mathrm{Uni}(Z{:}B|A) \\ +\mathrm{Red}(Z{:}A, B) + \mathrm{Syn}(Z{:}A, B)\end{aligned} \tag{1}$$

Here, $\mathrm{Uni}(Z{:}A|B)$ denotes the unique information about Z that is present only in A and not in B. E.g., *shopping preferences* (A) may provide unique information about *gender* (Z) that is not present in *address* (B). $\mathrm{Red}(Z{:}A, B)$ denotes the redundant information about Z that is present in both A and B. E.g., *zipcode* (A) and *county* (B) may provide redundant information about *race* (Z). The term $\mathrm{Syn}(Z{:}A, B)$ denotes the synergistic information not present in either A or B individually, but present jointly in (A, B), e.g., each individual digit of the *zipcode* may not have information about *race* but together they provide significant information. Before formally defining these terms, we provide an example.

Motivational Example. Let $Z{=}(Z_1, Z_2, Z_3)$ with each $Z_i{\sim}$ i.i.d. Bern(1/2). Let $A = (Z_1, Z_2, Z_3 \oplus N)$, $B = (Z_2, N)$, and $N \sim$ Bern(1/2) which is independent of Z. Here, $\mathrm{I}(Z; A, B) =$ 3 bits. The unique information about Z that is contained only in A and not in B is effectively in Z_1, and is given by $\mathrm{Uni}(Z{:}A|B) = \mathrm{I}(Z; Z_1) = 1$ bit. The redundant information about Z that is contained in both A and B is effectively in Z_2 and is given by $\mathrm{Red}(Z{:}A, B) = \mathrm{I}(Z; Z_2) = 1$ bit. Lastly, the synergistic information about Z that is not contained in either A or B alone, but is contained in both of them together is effectively in the tuple $(Z_3 \oplus N, N)$, and is given by $\mathrm{Syn}(Z{:}A, B){=}\mathrm{I}(Z; (Z_3 \oplus N, N)) = 1$ bit. This accounts for the 3 bits in $\mathrm{I}(Z; A, B)$.

We also note that defining any one of the PID terms suffices in obtaining the others. This is because of another relationship among the PID terms as follows [14]: $\mathrm{I}(Z; A) = \mathrm{Uni}(Z{:}A|B) + \mathrm{Red}(Z{:}A, B)$. Essentially $\mathrm{Red}(Z{:}A, B)$ is viewed as the sub-volume between $\mathrm{I}(Z; A)$ and $\mathrm{I}(Z; B)$ (see Fig. 1). Hence, $\mathrm{Red}(Z{:}A, B) = \mathrm{I}(Z; A) - \mathrm{Uni}(Z{:}A|B)$. Lastly, $\mathrm{Syn}(Z{:}A, B) = \mathrm{I}(Z; A, B) - \mathrm{Uni}(Z{:}A|B) - \mathrm{Uni}(Z{:}B|A) - \mathrm{Red}(Z{:}A, B)$ (can be obtained from (1) once both unique and redundant information has been defined).

The main results of our paper hold regardless of the specific definition of a given PID term. However, our experiments are based on the precise definition of $\mathrm{Uni}(Z{:}A|B)$ from [14].

Definition 1 (Unique Information [14])**.** *Let Δ be the set of all joint distributions on (Z, A, B) and Δ_p be the set of joint distributions with the same marginals on (Z, A) and (Z, B) as the true distribution, i.e., $\Delta_p = \{Q{\in}\Delta : \Pr_Q(Z = z, A = a) = \Pr(Z = z, A = a)$ and $\Pr_Q(Z = z, B = b) = \Pr(Z = z, B = b)\}$. Then,*

$$\mathrm{Uni}(Z{:}A|B) = \min_{Q \in \Delta_p} \mathrm{I}_Q(Z; A|B),$$

where $\mathrm{I}_Q(Z; A|B)$ is the conditional mutual information when (Z, A, B) have joint distribution Q and $\Pr_Q(\cdot)$ denotes the probability under Q.

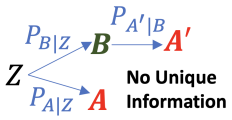

Fig. 2. Blackwell sufficiency of channel $P_{B|Z}$ with respect to $P_{A|Z}$ means A has no unique information about Z that is not in B.

Operational meaning of Unique Information from Blackwell sufficiency: Unique information is closely tethered to Blackwell Sufficiency [41] in statistical decision theory. The concept of Blackwell sufficiency [41] from statistical decision theory helps characterize if a random variable A is more informative than B about Z (also relates to stochastic degradation of channels [42], [43]). A channel $P_{B|Z}$ is Blackwell sufficient with respect to another channel $P_{A|Z}$ (also denoted as $B \geq_Z A$) if there exists a stochastic transformation $P_{A'|B}$ such that the effective channel from Z to A' is equivalent to the original channel from Z to A (see Fig. 2). The unique information $\mathrm{Uni}(Z{:}A|B)$ is 0 if and only if $P_{B|Z}$ is Blackwell sufficient with respect to $P_{A|Z}$ [14], [42]–[44]. Otherwise, $\mathrm{Uni}(Z{:}A|B) > 0$, and it is viewed as a departure from Blackwell sufficiency, i.e., *there exists a scenario where A*

gives something unique about Z that you can never get after degrading to B.

III. Partial Information Decomposition of the three Measures of Unfairness

We first introduce the information-theoretic quantification corresponding to the three definitions of fairness, namely, statistical parity, equalized odds, and predictive parity. Statistical parity (*independence*), requires the model prediction $\hat{Y}$ to be statistically independent of the sensitive attribute Z. Several measures have been proposed to quantify the gap from statistical parity [8], [45] (essentially dependence between $\hat{Y}$ and Z). In this work, we use the information-theoretic quantification of the statistical parity gap as defined next.

Definition 2 (Statistical Parity Gap)**.** *The statistical parity gap of a model f_θ with respect to Z is defined as $I(Z;\hat{Y})$, the mutual information between Z and $\hat{Y}$ (where $\hat{Y} = f_\theta(X)$).*

The concept of statistical parity has often been criticized for not considering the true labels. A perfect predictor $\hat{Y} = Y$ might not satisfy this criterion if Y is correlated to the sensitive attribute Z. Hence, the concept of equalized odds emerges as an alternative definition of fairness [7]. Equalized odds (*separation*) require the model's predictions $\hat{Y}$ to be independent of the sensitive attribute Z, conditioned on the true label Y, i.e., $Z \perp\!\!\!\perp \hat{Y}|Y$.

Definition 3 (Equalized Odds Gap)**.** *The equalized odds gap of a model f_θ with respect to Z is defined as $\mathrm{I}(Z;\hat{Y}|Y)$, the conditional mutual information between Z and $\hat{Y}$ given Y.*

Yet another vital fairness measure is predictive parity (*sufficiency*), which focuses on error parity among individuals given the same prediction [6]. Predictive parity requires the sensitive attribute Z to be independent of the true label Y conditioned on the model prediction $\hat{Y}$, i.e., $Z \perp\!\!\!\perp Y|\hat{Y}$.

Definition 4 (Predictive Parity Gap)**.** *The predictive parity gap of a model f_θ with respect to Z is defined as $\mathrm{I}(Z;Y|\hat{Y})$, the conditional mutual information between Z and Y given $\hat{Y}$.*

We leverage PID to derive exact relationships among the three measures of unfairness. We decompose the statistical parity gap $I(Z;\hat{Y})$, equalized odds gap $I(Z;\hat{Y}|Y)$, and predictive parity gap $I(Z;Y|\hat{Y})$ into *nonnegative* overlapping terms. The significance of this decomposition is that it highlights regions where these measures are in agreement and disagreement. Fig. 3 provides a pictorial illustration of the overlaps between these three measures of unfairness.

Proposition 1. *The statistical parity gap $\mathrm{I}(Z;\hat{Y})$, equalized odds gap $\mathrm{I}(Z;\hat{Y}|Y)$, and predictive parity gap $\mathrm{I}(Z;Y|\hat{Y})$ can be decomposed into nonnegative terms as follows:*

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y},Y). \tag{2}$$

$$\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Syn}(Z{:}\hat{Y},Y). \tag{3}$$

$$\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y). \tag{4}$$

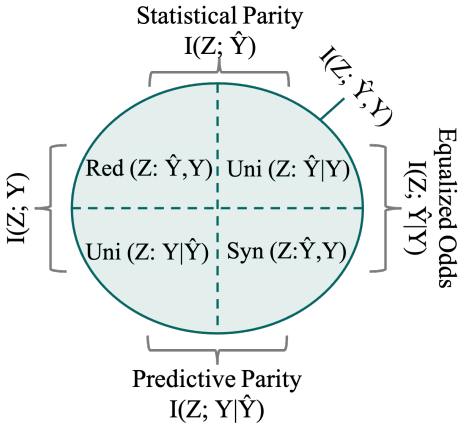

Fig. 3. Venn diagram showing the exact relationship between the various unfairness measures using PID: A critical observation is that all four PID terms are nonnegative. This enables us to derive several fundamental limits and tradeoffs among the unfairness measures, providing a nuanced understanding of when they agree and disagree.

The term $\mathrm{Uni}(Z{:}\hat{Y}|Y)$ quantifies the unique information about the sensitive attribute Z in the model prediction $\hat{Y}$ that is not there in the true label Y. $\mathrm{Uni}(Z{:}\hat{Y}|Y)$ is the common region between the statistical parity gap and the equalized odds gap, highlighting the region where they overlap. The term $\mathrm{Red}(Z{:}\hat{Y},Y)$ quantifies the information about sensitive attribute Z that is common between prediction $\hat{Y}$ and true label Y. $\mathrm{Red}(Z{:}\hat{Y},Y)$ contributes only to the statistical parity gap $\mathrm{I}(Z;\hat{Y})$ and not to any other measure of unfairness. The term $\mathrm{Syn}(Z{:}\hat{Y},Y)$ represents the synergistic information about sensitive attribute Z that is *not* present in either $\hat{Y}$ or Y individually but is present jointly in $(\hat{Y},S)$. $\mathrm{Syn}(Z{:}\hat{Y},Y)$ is the common region between equalized odds gap and predictive parity gap, highlighting their region of agreement. The unique information $\mathrm{Uni}(Z{:}Y|\hat{Y})$ contributes exclusively to the predictive parity gap $I(Z;Y|\hat{Y})$. This decomposition delineates the distinct regions where these unfairness measures overlap and diverge, offering a nuanced perspective on the interplay in machine learning models.

To better illustrate this decomposition, we now provide examples to understand each of these regions separately. Consider a hiring scenario featuring binary sensitive attributes and true labels i.e., $\hat{Y}, Z, Y \in \{0,1\}$ with $Z\sim$ Bern(1/2).

Example 1 (Pure Uniqueness to Model Prediction)**.** *Let $\hat{Y} = Z$ and $Z \perp\!\!\!\perp Y$ (an equal base rate for privileged and unprivileged groups). Suppose, the model only approves privileged candidates ($Z = 1$) but rejects the unprivileged ($Z = 0$). This model violates both statistical parity and equalized odds, i.e., $\mathrm{I}(Z;\hat{Y}) = \mathrm{I}(Z;\hat{Y}|Y) = 1$. This model satisfies predictive parity criterion, $\mathrm{I}(Z;Y|\hat{Y}) = 0$. This is a case of purely unique information in the model prediction that is not in the true label since all the information about Z is derived exclusively from the model predictions; the true label Y does not correlate with Z. Here, $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 1$, $\mathrm{Red}(Z{:}\hat{Y},Y) = 0$, $\mathrm{Syn}(Z{:}\hat{Y},Y) = 0$, and $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0$.*

Example 2 (Pure Redundancy)**.** *Let $\hat{Y} = Y$ and $Y = Z$ with probability 0.9. There is a correlation between the true label Y and protected attribute Z, but this model has perfect accuracy. Such a model satisfies equalized odds and predictive parity criterion, i.e., $\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{I}(Z;Y|\hat{Y}) = 0$. However, the model fails to satisfy statistical parity since $\mathrm{I}(Z;\hat{Y}) = 0.53$. This is a case of purely redundant information since the information about Z is entirely common between both $\hat{Y}$ and Y. Here, $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0$, $\mathrm{Red}(Z{:}\hat{Y},Y) = 0.53$, $\mathrm{Syn}(Z{:}\hat{Y},Y) = 0$, and $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0$.*

Example 3 (Pure Synergy)**.** *Let $\hat{Y} = Z$ XNOR Y and $Z \perp\!\!\!\perp Y$. The model approves candidates from the privileged group ($Z = 1$) with true label $Y = 1$, and also from the unprivileged group ($Z = 0$) with $Y = 0$. On the other hand, it rejects candidates from the unprivileged group ($Z = 0$) with true label $Y = 1$, and the privileged group ($Z = 1$) with true label $Y = 0$. Such a model violates equalized odds (and predictive parity) as it singularly prefers one group within each true label class. Thus, $\mathrm{I}(Z;\hat{Y}|Y) = 1$, and $\mathrm{I}(Z;Y|\hat{Y}) = 1$. However, it achieves statistical parity since it maintains an equal approval rate for both privileged and unprivileged groups with $\mathrm{I}(Z;\hat{Y}) = 0$. This is a case of synergistic information about Z that is not observable in either $\hat{Y}$ or Y individually but is present jointly in $Y,\hat{Y}$. Here, $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0$, $\mathrm{Red}(Z{:}\hat{Y},Y) = 0$, $\mathrm{Syn}(Z{:}\hat{Y},Y) = 1$, and $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0$.*

Example 4 (Pure Uniqueness to True Label)**.** *Let $Y = Z$ with probability 0.9 and $Z \perp\!\!\!\perp \hat{Y}$. The true label Y is highly correlated to sensitive attribute Z, but the model prediction $\hat{Y}$ is independent of sensitive attribute Z. This model violates predictive parity ($\mathrm{I}(Z;Y|\hat{Y}) = 0.53$) but satisfies statistical parity and equalized odds ($\mathrm{I}(Z;\hat{Y}) = \mathrm{I}(Z;\hat{Y}|Y) = 0$). This is a case of unique information about sensitive attributes in the true label that is not in the model prediction. Here, $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0$, $\mathrm{Red}(Z{:}\hat{Y},Y) = 0$, $\mathrm{Syn}(Z{:}\hat{Y},Y) = 0$, and $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0.53$.*

These examples demonstrate scenarios of pure uniqueness, redundancy, and synergy to help us understand the decomposition. PID serves as a tool to highlight regions of agreement and disagreement between these fairness definitions. In contrast, traditional fairness metrics lack the granularity to capture these nuanced interactions, making PID an essential asset for a more comprehensive understanding and mitigation of disparities.

We can go beyond the impossibility between the three fairness definitions and further analyze their interrelationships.

Theorem 1 (Revisiting Impossibility)**.** *If $\mathrm{I}(Z;\hat{Y},Y) > 0$, at least one of the PID terms, namely, $\mathrm{Uni}(Z{:}\hat{Y}|Y)$, $\mathrm{Red}(Z{:}\hat{Y},Y)$, $\mathrm{Syn}(Z{:}\hat{Y},Y)$, or $\mathrm{Uni}(Z{:}Y|\hat{Y})$ will be nonnegative. Hence, at least one of the fairness measures, namely, the Statistical Parity Gap ($\mathrm{I}(Z;\hat{Y})$), Equalized Odds Gap ($\mathrm{I}(Z;\hat{Y}|Y)$), or Predictive Parity Gap ($\mathrm{I}(Z;Y|\hat{Y})$) will be nonzero. Conversely, all these unfairness measures will be zero if and only if $\mathrm{I}(Z;\hat{Y},Y) = 0$.*

Proof Sketch: The proof relies on the nonnegativity of each of the PID terms (also recall Fig. 3). PID of $\mathrm{I}(Z;\hat{Y},Y)$ is expressed as $\mathrm{I}(Z;\hat{Y},Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y},Y) + \mathrm{Syn}(Z{:}\hat{Y},Y)$. Since each component in this decomposition is nonnegative, the presence of mutual information ($\mathrm{I}(Z;\hat{Y},Y) > 0$) implies that at least one of these terms has a nonzero contribution. According to Proposition 1, each of these PID terms influences at least one unfairness measure. Therefore, the nonnegativity of any one of these terms results in at least one of the unfairness measures being nonzero. □

This is a general result from which one can also derive the impossibility of the three fairness definitions under specific conditions. Our next result examines the unfairness measures only when $\mathrm{I}(Z;Y) > 0$. It is important to note that $\mathrm{I}(Z;Y)$ is an inherent characteristic of the dataset alone and hence it is independent of the model predictions.

Theorem 2 (Dataset Dependent Relationships)**.** *If $\mathrm{I}(Z;Y) > 0$, either the Statistical Parity Gap $\mathrm{I}(Z;\hat{Y})$ or the Predictive Parity Gap $\mathrm{I}(Z;Y|\hat{Y})$ must be greater than zero.*

Proof Sketch: The proof relies on demonstrating that the mutual information between Z and Y can be expressed as:

$$\mathrm{I}(Z;Y) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}Y,\hat{Y}). \tag{5}$$

Though, the PID terms $\mathrm{Uni}(Z{:}Y|\hat{Y})$ and $\mathrm{Red}(Z{:}Y,\hat{Y})$ may vary based on the model chosen, their sum remains constant, reflecting the fixed nature of the mutual information between Z and Y in the dataset. Notably, $\mathrm{Uni}(Z{:}Y|\hat{Y})$ contributes to the predictive parity gap, and $\mathrm{Red}(Z{:}Y,\hat{Y})$ contributes to the statistical parity gap (recall Fig. 3). □

IV. Tradeoffs Between Unfairness Measures

In this section, we delineate the fundamental limits and tradeoffs between various unfairness measures. Our findings underscore the intricate and sometimes conflicting nature of different fairness objectives in algorithmic decision-making. Examining fairness through the lens of PID uncovers the nuanced interplay between different unfairness measures.

We explore scenarios where models are trained with a focus on achieving any one specific fairness criterion and analyze its implications on the other two fairness notions. This applies to models that have been trained to achieve fairness either through in-processing techniques, such as adding fairness regularizers to the loss function, or through post-processing methods that adjust model outputs after training.

Theorem 3. *If Statistical Parity is satisfied, i.e., $\mathrm{I}(Z;\hat{Y}) = 0$, then the Predictive Parity Gap is greater than the Equalized Odds Gap, i.e., $\mathrm{I}(Z;Y|\hat{Y}) \geq \mathrm{I}(Z;\hat{Y}|Y)$. Additionally, if the dataset is such that $\mathrm{I}(Z;Y) = 0$, then Predictive Parity and Equalized Odds are equivalent, i.e., $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{I}(Z;\hat{Y}|Y)$.*

Proof Sketch: We refer to Fig. 3 for an intuitive understanding of the proof. Given that Statistical Parity is zero, we have $\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y},Y) = 0$.

TABLE I
RESULTS OF REGULARIZERS ON DIFFERENT MEASURES OF UNFAIRNESS

Regularizers	Equalized Odds $\mathrm{I}(Z;\hat{Y}\|Y)$			
	Statistical Parity $\mathrm{I}(Z;\hat{Y})$		Predictive Parity $\mathrm{I}(Z;Y\|\hat{Y})$	
	$\mathrm{Red}(Z{:}\hat{Y},Y)$	$\mathrm{Uni}(Z{:}\hat{Y}\|Y)$	$\mathrm{Syn}(Z{:}\hat{Y},Y)$	$\mathrm{Uni}(Z{:}Y\|\hat{Y})$
SP	0.012	0.000	0.001	0.024
PP	0.026	0.007	0.008	0.011
EO	0.011	0.000	0.001	0.026
EO, PP	0.000	0.000	0.000	0.037
SP, PP	0.000	0.000	0.000	0.037
SP, EO	0.008	0.000	0.000	0.028
SP, EO, PP	0.000	0.000	0.000	0.037

Since all PID terms are non-negative, it follows that individually $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0$ and $\mathrm{Red}(Z{:}\hat{Y},Y) = 0$. Consequently, the Equalized Odds gap simplifies to $\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{Syn}(Z{:}\hat{Y},Y)$. On the other hand, the Predictive Parity gap is $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y)$. Since all PID terms are nonnegative, it follows that $\mathrm{I}(Z;Y|\hat{Y}) \geq \mathrm{I}(Z;\hat{Y}|Y)$.

Furthermore, when $\mathrm{I}(Z;Y) = 0$, it results in $\mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y},Y) = 0$, leading to each of those individual terms being zero, i.e., $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0$ and $\mathrm{Red}(Z{:}\hat{Y},Y) = 0$. Therefore, $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{I}(Z;\hat{Y}|Y)$. □

Similar to Theorem 3, one can also derive the relationship between the statistical parity gap and equalized odds gap when predictive parity is satisfied.

Theorem 4. *If Predictive Parity is satisfied, i.e.,* $\mathrm{I}(Z;Y|\hat{Y}) = 0$*, then the Statistical Parity Gap is greater than the Equalized Odds Gap, i.e.,* $\mathrm{I}(Z;\hat{Y}) \geq \mathrm{I}(Z;\hat{Y}|Y)$*. Additionally, if the dataset is such that* $\mathrm{I}(Z;Y) = 0$*, then Statistical Parity and Equalized Odds are equal, i.e.,* $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{I}(Z;\hat{Y}|Y)$.

Theorem 3 & 4 demonstrate scenarios where one unfairness measure dominates another and are in agreement, now we provide a third scenario where two measures of unfairness will be in disagreement.

Theorem 5. *If Equalized Odds is satisfied, i.e.,* $\mathrm{I}(Z;\hat{Y}|Y) = 0$ *and* $\mathrm{I}(Z;Y) > 0$*, an inverse relationship (tradeoff) exists between Statistical Parity and Predictive Parity, i.e.,* $\mathrm{I}(Z;\hat{Y}) = \mathrm{I}(Z;Y) - \mathrm{I}(Z;Y|\hat{Y})$*. Thus, increasing one leads to a decrease in the other, and vice versa.*

Proof Sketch Given that Equalized Odds is met, we have $\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Syn}(Z{:}\hat{Y},Y) = 0$. Consequently, from nonnegativity, both the terms $\mathrm{Uni}(Z{:}\hat{Y}|Y)$ and $\mathrm{Syn}(Z{:}\hat{Y},Y)$ are 0. Statistical Parity gap simplifies to $\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y},Y) = \mathrm{Red}(Z{:}\hat{Y},Y)$, and the Predictive Parity gap is expressed as $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{Uni}(Z{:}Y|\hat{Y})$. Hence, $\mathrm{I}(Z;Y) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y},Y) = \mathrm{I}(Z;\hat{Y}) + \mathrm{I}(Z;Y|\hat{Y})$. Since, $\mathrm{I}(Z;Y)$ is fixed for a dataset, an increase in the statistical parity gap leads to a decrease in the predictive parity gap, and vice versa. □

V. EXPERIMENTAL DEMONSTRATIONS

In this section, we provide an experimental demonstration on the Adult dataset [16] to validate our theoretical findings. The classification task for this dataset involves predicting whether an individual's income exceeds 50K per year, using features such as occupation, marital status, and education. We use *gender* as a sensitive attribute.

We train a neural network consisting of a sequence of layers: the input layer is followed by three hidden layers, each with 32 units and ReLU activation, and concludes with a single output layer using a sigmoid activation function. Training is conducted using a batch size of 512, and the Adam optimizer with a learning rate of 0.01. We apply various fairness regularizers and measure the unfairness as well as their decomposition (results are summarized in Table.I). We use the *dit* package [46] for PID computation and *FairTorch* [47] for fairness regularizer implementation.

A key observation in our analysis is that $\mathrm{I}(Z;Y)$ consistently measures 0.037 using the Adult dataset. This mass does not decrease across various models since it only depends on the dataset. The PID terms in $\mathrm{I}(Z;Y)$, i.e., $\mathrm{Uni}(Z{:}Y|\hat{Y})$ and $\mathrm{Red}(Z{:}\hat{Y},Y)$ contribute to either predictive parity or statistical parity gap. Also, when statistical parity is achieved (scenario with SP regularizer), the predictive parity gap is greater than the equalized odds gap. Also due to the impossibility of attaining zero unfairness with all the measures (see scenario with SP, EO, and PP regularizers), the mass typically moves to $\mathrm{Uni}(Z{:}Y|\hat{Y})$, contributing to the predictive parity.

VI. CONCLUSION

By introducing this unifying framework, we provide a tool for gaining a more nuanced understanding of the interplay between different unfairness measures, thereby enhancing the decision-making process in the deployment of fair ML systems. Our work holds broader implications in fields such as algorithmic fairness auditing [17], explainability [48], policy regulation [1], where it can significantly contribute to the evaluation and understanding of unfairness in ML models. This work not only furthers the theoretical discourse but would also have significant societal implications, guiding the trajectory toward more responsible and equitable machine learning in high-stakes settings.

REFERENCES

[1] The White House, "Blueprint for an ai bill of rights: Making automated systems work for the american people," https://www.whitehouse.gov/ostp/ai-bill-of-rights/, 2022, accessed: [30 Jan, 2024].

[2] K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY: Independently Published, 2021.

[3] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 643–650.

[4] S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover, "Fairness under feature exemptions: Counterfactual and observational measures," *IEEE Transactions on Information Theory*, vol. 67, no. 10, pp. 6675–6710, 2021.

[5] S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover, "An information-theoretic quantification of discrimination with exempt features," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[6] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surv.*, vol. 55, no. 3, feb 2022.

[7] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, 2016.

[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[9] A. L. Washington, "How to argue with an algorithm: Lessons from the compas-propublica debate," *Colo. Tech. LJ*, vol. 17, p. 131, 2018.

[10] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, p. 671, 2016.

[11] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, "A clarification of the nuances in the fairness metrics landscape," *Scientific Reports*, vol. 12, no. 1, p. 4209, 2022.

[12] S. Dutta, D. Wei, H. Yueksel, P. Y. Chen, S. Liu, and K. R. Varshney, "Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing," in *International Conference on Machine Learning (ICML)*, 2020, pp. 2803–2813.

[13] J. S. Kim, J. Chen, and A. Talwalkar, "Fact: A diagnostic for group fairness trade-offs," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5264–5274.

[14] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," *Entropy*, vol. 16, no. 4, pp. 2161–2183, 2014.

[15] A. Ghassami, S. Khodadadian, and N. Kiyavash, "Fairness in supervised learning: An information theoretic approach," in *2018 IEEE international symposium on information theory (ISIT)*, 2018, pp. 176–180.

[16] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[17] T. Yan and C. Zhang, "Active fairness auditing," in *International Conference on Machine Learning*. PMLR, 2022, pp. 24 929–24 962.

[18] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer, 2012, pp. 35–50.

[19] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," *Advances in neural information processing systems*, vol. 30, 2017.

[20] J. Cho, G. Hwang, and C. Suh, "A fair classifier using mutual information," in *2020 IEEE international symposium on information theory (ISIT)*. IEEE, 2020, pp. 2521–2526.

[21] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn, "Renyi fair inference," *arXiv preprint arXiv:1906.12005*, 2019.

[22] V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki, "Fairness-aware neural reyni minimization for continuous features," *arXiv preprint arXiv:1911.04929*, 2019.

[23] H. Wang, H. Hsu, M. Diaz, and F. P. Calmon, "The impact of split classifiers on group fairness," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE Press, 2021, p. 3179–3184.

[24] S. Galhotra, K. Shanmugam, P. Sattigeri, and K. R. Varshney, "Causal feature selection for algorithmic fairness," in *Proceedings of the 2022 International Conference on Management of Data*, 2022, pp. 276–285.

[25] W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P. Michalak, S. Asoodeh, and F. Calmon, "Beyond adult and compas: Fair multi-class prediction via information projection," in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 2022, pp. 38 747–38 760.

[26] P. Kairouz, J. Liao, C. Huang, M. Vyas, M. Welfert, and L. Sankar, "Generating fair universal representations using adversarial models," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1970–1985, 2022.

[27] F. Hamman, J. Chen, and S. Dutta, "Can querying for bias leak protected attributes? achieving privacy with smooth sensitivity," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1358–1368.

[28] L. Chu, L. Wang, Y. Dong, J. Pei, Z. Zhou, and Y. Zhang, "Fedfair: Training fair models in cross-silo federated learning," *arXiv preprint arXiv:2109.05662*, 2021.

[29] B. Rodríguez-Gálvez, F. Granqvist, R. van Dalen, and M. Seigel, "Enforcing fairness in private federated learning via the modified method of differential multipliers," *arXiv preprint arXiv:2109.08604*, 2021.

[30] D. Y. Zhang, Z. Kou, and D. Wang, "Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 1051–1060.

[31] H. Wang, L. He, R. Gao, and F. Calmon, "Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[32] C. X. Long, H. Hsu, W. Alghamdi, and F. Calmon, "Individual arbitrariness and group fairness," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=nzkWhoXUpv

[33] S. Dutta and F. Hamman, "A review of partial information decomposition in algorithmic fairness and explainability," *Entropy*, vol. 25, no. 5, p. 795, 2023.

[34] F. Hamman and S. Dutta, "Demystifying local and global fairness trade-offs in federated learning using partial information decomposition," *International Conference on Learning Representations (ICLR)*, 2024.

[35] D. A. Ehrlich, A. C. Schneider, M. Wibral, V. Priesemann, and A. Makkeh, "Partial information decomposition reveals the structure of neural representations," *arXiv preprint arXiv:2209.10438*, 2022.

[36] T. M. Tax, P. A. Mediano, and M. Shanahan, "The partial information decomposition of generative neural network models," *Entropy*, vol. 19, no. 9, p. 474, 2017.

[37] P. P. Liang, Y. Cheng, X. Fan, C. K. Ling, S. Nie, R. Chen, Z. Deng, F. Mahmood, R. Salakhutdinov, and L.-P. Morency, "Quantifying & modeling feature interactions: An information decomposition framework," *arXiv preprint arXiv:2302.12247*, 2023.

[38] P. Wollstadt, S. Schmitt, and M. Wibral, "A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition." *J. Mach. Learn. Res.*, vol. 24, pp. 131–1, 2023.

[39] S. Mohamadi, G. Doretto, and D. A. Adjeroh, "More synergy, less redundancy: Exploiting joint mutual information for self-supervised learning," *arXiv preprint arXiv:2307.00651*, 2023.

[40] A. Pakman, A. Nejatbakhsh, D. Gilboa, A. Makkeh, L. Mazzucato, M. Wibral, and E. Schneidman, "Estimating the unique information of continuous variables," *Advances in neural information processing systems*, vol. 34, pp. 20 295–20 307, 2021.

[41] D. Blackwell, "Equivalent comparisons of experiments," *The annals of mathematical statistics*, pp. 265–272, 1953.

[42] P. K. Banerjee, E. Olbrich, J. Jost, and J. Rauh, "Unique informations and deficiencies," in *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2018, pp. 32–38.

[43] P. Venkatesh, K. Gurushankar, and G. Schamberg, "Capturing and interpreting unique information," in *IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 2631–2636.

[44] P. Venkatesh and G. Schamberg, "Partial information decomposition via deficiency for multivariate gaussians," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 2892–2897.

[45] Y. Shi, H. Yu, and C. Leung, "A survey of fairness-aware federated learning," *arXiv preprint arXiv:2111.01872*, 2021.

[46] R. G. James, C. J. Ellison, and J. P. Crutchfield, "dit: a Python package for discrete information theory," *The Journal of Open Source Software*, vol. 3, no. 25, p. 738, 2018.

[47] M. S. Akihiko Fukuchi, Yoko Yabe, "Fairtorch," https://github.com/wbawakate/fairtorch, 2021.

[48] J. Zhou, F. Chen, and A. Holzinger, "Towards explainability for ai fairness," in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 2020, pp. 375–386.