Chapter 17

Relationships between Differential Privacy and Algorithmic Fairness

By Rachel Cummings

DOI: 10.1561/9781638284772.ch17

Published in *Differential Privacy in Artificial Intelligence: From Theory to Practice* by Ferdinando Fioretto and Pascal Van Hentenryck (eds.). 2025. ISBN 978-1-63828-476-5. E-ISBN 978-1-63828-477-2.

Suggested citation: Rachel Cummings. 2025. "Relationships between Differential Privacy and Algorithmic Fairness" in *Differential Privacy in Artificial Intelligence: From Theory to Practice*. Edited by Ferdinando Fioretto and Pascal Van Hentenryck. pp. 557–590. Now Publishers. DOI: 10.1561/9781638284772.ch17.

now
the essence of knowledge

17.1 Introduction

Recent years have found growing interest in the overlap between the fields of differential privacy and algorithmic fairness. This includes both the question of when privacy and fairness can be achieved together in a learning system, and questions about the underlying technical relationship between privacy and fairness. While these questions are natural and simple to pose, they are substantially less straightforward to answer.

One major challenge is that there does not exist a universal definition of algorithmic fairness. This is in contrast to differential privacy, which is a single mathematical formalization of privacy[1]. At a high level, all fairness definitions take the form: "give similar treatment to people who deserve to be treated similarly," although there are many alternative formalizations of "similar treatment" and "people who deserve to be treated similarly." For example, should decision-making rules only consider an individual's observable, non-sensitive attributes (e.g., school admission based only on standardized test scores), or should it also consider protected

1. Although importantly, DP is not the only privacy definition one may ever wish to consider

attributes for the sake of de-biasing other measurements or as recompense for historical inequities (e.g., affirmative action)? To further complicate matters, analysts often must choose the set of attributes that deserve protection in their dataset, either because this set is not defined exogenously in the law, or because there are *proxies* that correlate with the sensitive attribute and can therefore be disclosive (e.g., shampoo choice correlates with race and gender). Additionally, the guarantee of "similar treatment" can be formalized with respect to many different performance metrics, such as precision or recall of a binary classifier, probability of a correct classification, closeness in distribution of outcomes, and more.

The fairness literature for the most part can be divided into three different classes of definitions:

1. **individual fairness** requires that "similar individuals are treated similarly,"
2. **group fairness** requires "fairness with respect to protected group membership," and
3. **multi-group fairness** requires "fairness with respect to membership in many, potentially overlapping groups".

These three classes are respectively explored in Sections 17.2, 17.3, and 17.4; each section presents formal fairness definitions, algorithms that achieve the desired fairness notions, and the relationship between the fairness notion and differential privacy.

In this chapter, we primarily consider the task of *binary classification*, where each individual i has some observable attributes X_i and a binary label $Y_i \in \{0, 1\}$; where appropriate, they will also have a protected attribute A_i. For individual fairness in Section 17.2, the fair algorithm only observes X_is, and must assign an outcome $\hat{Y}_i$ to each individual; for group and multi-group fairness in Sections 17.3 and 17.4, the algorithm takes in a training set of n observations containing (X_i, Y_i) pairs (and A_i in Section 17.3), and must produce a binary classifier for use on future observations that maps X_i to a predicted outcome $\hat{Y}_i$. In all settings, the mapping from X_i to $\hat{Y}_i$ must respect the fairness constraint.

As running examples of fair learning tasks throughout this chapter, we will use (1) school admissions and (2) evaluating loan applications. In each these tasks, binary decisions are made about individuals (i.e., acceptance/rejection at a school; approval/denial of a loan) based on that individual's submitted features X. In both of these settings, there are certain social, moral, and legal fairness obligations that should be respected in the decision-making processes. While we use these two examples for concreteness, we also emphasize that these are far from the only application domains where fairness requirements are relevant.

When combining differential privacy and algorithmic fairness in a machine learning system, a key observation is that these two constraints apply to different

components of the learning pipeline. The classifier must be learned *privately* with respect to the training data, in order to not leak sensitive information about the training data when the classifier is applied in the future. The resulting classifier must also be *fair* when applied to new (test) data. That is, privacy is required in training, and fairness is required when testing. Of course, these requirements must both be considered together during training, to ensure that the privately learned classifier will be fair. Section 17.5 explores problems that may arise when fairness is *not* explicitly considered during training.

Before delving into the material, it is important to clarify that this chapter is *not* intended to be a comprehensive survey of the fairness literature, which is itself well-developed and rapidly evolving. Instead this chapter focuses only the parts of the fairness literature that are most relevant for differential privacy. The interested reader is referred elsewhere for a more comprehensive overview of the fairness field.

17.2 Individual Fairness

Individual fairness, first introduced in Dwork et al. [Dwo+12], is arguably the most natural of all the fairness notions considered in this chapter. At a high level, it requires that *similar individuals should be treated similarly*. In the language of our canonical running examples, this requires that students with similar academic aptitudes should be admitted to similar schools, and that loan applicants with similar likelihood of repaying their loans should receive loans with similar conditions. This adheres nicely to the primary high-level principle of differential privacy, that *similar datasets should be treated similarly* under differentially private algorithms. As we will see, this enables the tools of differential privacy to be brought to bear in a very organic fashion to achieve individual fairness—essentially by treating individuals with their many attributes as databases with their many entries.

The algorithmic approach of Dwork et al. [Dwo+12] described in Section 17.2.1 considers a classification setting, and relies on the existence of metrics over both individuals and outcomes to measure similarity for each. The individually fair classifiers they construct are mappings from individuals to outcomes that approximately preserve distances with respect to these metrics. They show that the use of differentially private algorithms is one such method for achieving this fairness goal, which is presented concretely here in Section 17.2.2.

As we will also see, this idealized fairness notion is unfortunately difficult to achieve in practice, as it requires a perfect metric of "distance over individuals" to measure similarity. For most—if not all—practical applications where fairness is desired, no such metric exists, and no such measurement of individuals can be made. What is a perfect measure of student aptitude or of willingness to repay a

loan? We can use SAT scores and credit scores, respectively, but these are known to be imperfect measures that can be undesirably correlated with protected attributes such as race, gender, and family background, and using such imperfect measures risks further embedding existing societal biases. Recent work of Ilvento [Ilv20], Rothblum and Yona [RY18], and Jung et al. [Jun+20] have all explored more practical and query-efficient ways of learning a metric that can be used in the algorithmic individual fairness framework of Dwork et al. [Dwo+12]; these are discussed in Section 17.2.4.

17.2.1 Setting and Results

Consider a classification task over individuals (e.g., a bank must approve or deny loans based on applications). Let $\mathcal{X}$ be the set of individuals, and let $\mathcal{Y}$ be the set of outcomes (e.g., $\mathcal{Y} = \{0, 1\}$ for binary classification, as in our examples). The classification task is to predict an individual's $\hat{Y} \in \mathcal{Y}$ given their observed attributes $X \in \mathcal{X}$. Assume there exists a metric on individuals $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which takes as input two individuals and outputs a "similarity score". That is, $d(X_1, X_2)$ quantifies the similarity of individuals $X_1, X_2 \in \mathcal{X}$.

We consider random classifiers $M : \mathcal{X} \to \Delta(\mathcal{Y})$, which map individuals to distributions over outcomes. That is, given an individual $X \in \mathcal{X}$, the classification mechanism will choose an outcome according to the distribution $M(X)$. Our fairness goal is for this mechanism to map similar individuals to similar distributions, so we also require a distance measure D on distributions to quantify similarity, for example D_∞ or D_{TV}, defined below.

Definition 17.1 (Total variation distance, D_{TV})**.** *Let P, Q be two probability measures over a finite domain A. The statistical distance or total variation distance between P and Q is:*

$$D_{TV}(P, Q) = \frac{1}{2} \sum_{Y \in \mathcal{Y}} |P(Y) - Q(Y)|.$$

Note that $D_{TV}(P, Q) \in [0, 1]$ for all distributions P, Q. If P and Q are similar, then $D_{TV}(P, Q)$ will be close to 0, and if P and Q are very different, then $D_{TV}(P, Q)$ will be close to 1.

Definition 17.2 (D_∞ distance)**.** *Let P, Q be two probability measures over a finite domain A. The D_∞ distance between P and Q is:*

$$D_\infty(P, Q) = \sup_{Y \in \mathcal{Y}} \log \left(\max \left\{ \frac{P(Y)}{Q(Y)}, \frac{Q(Y)}{P(Y)} \right\} \right).$$

Note that $D_\infty(P, Q) \in [0, \infty)$ for all distributions P, Q. If P and Q are similar, then $D_\infty(P, Q) \ll 1$. If P and Q are very different, then $D_\infty(P, Q) \gg 1$.

We formalize the individual fairness constraint by requiring M to be a Lipschitz mapping with respect to the metrics D and d.

Definition 17.3 (Lipschitz mapping)**.** *A mapping $M : \mathcal{X} \to \Delta(\mathcal{Y})$ is (D, d)-Lipschitz if for all $X_1, X_2 \in \mathcal{X}$,*

$$D(M(X_1), M(X_2)) \leq d(X_1, X_2).$$

Definition 17.4 (Individually fair classifier [Dwo+12])**.** *A classifier $M : \mathcal{X} \to \Delta(\mathcal{Y})$ is* individually fair *with respect to metrics $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $D : \Delta(\mathcal{Y}) \times \Delta(\mathcal{Y}) \to \mathbb{R}$ if it is (D, d)-Lipschitz.*

Note that, e.g., constant classifiers are trivially fair. Thus we want an individually fair classifier that also satisfies a notion of utility. Let $L : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a loss function that measures the utility of assigning a specific outcome to a specific individual. In the example of loan applications, L may capture the expected probability that an applicant will default on their loan, or the (negative) expected revenue from awarding a loan to this applicant.

The goal of individual fairness is described in the following optimization problem:

Find a mapping M that minimizes expected loss subject to the Lipschitz condition.

This optimization problem can be stated mathematically as follows:

$$\begin{aligned} &\min_M \mathop{\mathbb{E}}_{X \sim \mathcal{X}} \mathop{\mathbb{E}}_{\hat{Y} \sim M(X)} L(X, \hat{Y}) \\ &\text{s.t. } D(M(X_1), M(X_2)) \leq d(X_1, X_2), \quad \forall X_1, X_2 \in \mathcal{X} \\ &\quad M(X) \in \Delta(\mathcal{Y}), \quad \forall X \in \mathcal{X}. \end{aligned} \tag{17.1}$$

Assuming oracle access to $d(X_1, X_2)$ and $L(X, \hat{Y})$, and when D is D_∞ or D_{TV}, this problem can be written as a linear program and solved efficiently.

Theorem 17.5 ([Dwo+12])**.** *When D is either D_{TV} or D_∞, the optimization problem described in (17.1) can be solved with an LP of size poly$(|\mathcal{X}|, |\mathcal{Y}|)$.*

To see this, we can write the classifier M as a collection of distributions $\mu_X = M(X) \in \Delta(\mathcal{Y})$, which can each in turn be described by a collection of decision variables $\mu_X(\hat{Y}) = \mathbf{Pr}[M(X) = \hat{Y}]$, for all $X \in \mathcal{X}$ and $\hat{Y} \in \mathcal{Y}$. Then it is easy to see that the objective of (17.1) is linear in the $|\mathcal{X}||\mathcal{Y}|$ decision variables, and that the second constraint can be written as a collection of $|\mathcal{X}|$ linear constraints.

When $D = D_{TV}$, the first constraint of (17.1) can be re-written as $\frac{1}{2}\sum_{\hat{Y}\in\mathcal{Y}}(\mu_{X_1}(\hat{Y}) - \mu_{X_2}(\hat{Y})) \leq d(X_1, X_2)$ and $\frac{1}{2}\sum_{\hat{Y}\in\mathcal{Y}}(\mu_{X_2}(\hat{Y}) - \mu_{X_1}(\hat{Y})) \leq d(X_1, X_2)$, which requires $2|\mathcal{X}|^2$ linear constraints, given explicit access to the metric d.

When $D = D_\infty$, we can re-write the first constraint as:

$$\log \frac{\mu_{X_1}(\hat{Y})}{\mu_{X_2}(\hat{Y})} \leq d(X_1, X_2) \iff \mu_{X_1}(\hat{Y}) \leq e^{d(X_1,X_2)}\mu_{X_2}(\hat{Y}),$$
$$\forall X_1, X_2 \in \mathcal{X},\ \forall \hat{Y} \in \mathcal{Y}.$$

Given explicit access to the metric d, this requires only $|\mathcal{X}|^2|\mathcal{Y}|$ linear constraints.

Other choices of metric D can also be plugged into this framework, and the same optimization problem of (17.1) still applies. The computational efficiency of solving this optimization problem will depend on on properties of the chosen metric, and how well the resulting instantiation of (17.1) can be solved using commercial solvers.

17.2.2 Relationship to Differential Privacy

The connection between individual fairness and differential privacy is immediate from the problem description. Individual fairness considers a mechanism that takes in an *individual* and outputs an outcome sampled from a distribution. Individual fairness requires that the distribution over outputs is similar when the the mechanism is run on *similar individuals*. Differential privacy considers a mechanism which takes in a *database* and outputs an outcome sampled from a distribution. Differential privacy requires that the distribution over outputs is similar when the mechanism is run on *similar databases*.

Using the language of differential privacy with the notation of individual fairness, consider a database $X \in \mathcal{X}$ and an analyst who wishes to privately answer a query $f : \mathcal{X} \to \mathcal{Y}$. The analyst will respond using a randomized mechanism $M : \mathcal{X} \to \Delta(\mathcal{Y})$. This mechanism will be ε-differentially private if and only if M is (D_∞, d)-Lipschitz for distance metric $d(X_1, X_2) = \varepsilon\|X_1 - X_2\|_1$. The utility loss function for producing answer $\hat{Y} \in \mathcal{Y}$ on database X is $L(X, \hat{Y}) = d_f(f(X), \hat{Y})$, which should capture the domain-specific accuracy notion. A common loss function for real-valued queries would be additive error: $d_f(f(X), \hat{Y}) = |f(X) - \hat{Y}|$.

In short, differential privacy can be seen as an instantiation of individual fairness, where databases are the individuals, and the neighboring relationship defines similarity. Hence many of the algorithmic tools developed for differential privacy can also be immediately applied to achieve individual fairness with the D_∞ metric over outcome distributions.

17.2.3 Pros and Cons of Individual Fairness

Pros

The biggest strength of individual fairness is that it is a very natural and strong fairness notion, which formalizes exactly what we would like our fairness notions to capture: people should be considered as individuals and receive treated in the way that they deserve, based on problem-specific merit-measures rather than irrelevant attributes; e.g., admit students to schools based on aptitude, not based on parents' income. The fairness constraint holds *for all* individuals, not just for some.

The natural connection to differential privacy allows existing DP tools to be used as individually fair mechanisms. Additionally, the general LP framework of (17.1) allows individually fair classifiers to efficiently computable if $|\mathcal{X}|$ is reasonably sized.

Finally, since fairness is considered at the individual level, algorithm designers do not need to pre-specify a class of protected groups as is required for group fairness (Section 17.3). This avoids problems of proxy variables which are correlated with the protected attribute (e.g., hair length as a proxy for gender). It also ensures fairness within groups, since "better" individuals receive "better" outcomes, independent of group membership. As we will see in Section 17.3, these are major challenges when applying group fairness.

Cons

The biggest weakness of individual fairness is that it requires complete knowledge of a perfect distance metric d over individuals. This is nearly impossible to achieve in practice, where the existing individual-level metrics are known to be imperfect and embed existing biases (e.g., standardized test scores and credit scores). Recent developments to address this challenge are discussed in Section 17.2.4.

Computationally, the population size $|\mathcal{X}|$ is not reasonably sized in many practical applications, and even $poly(|\mathcal{X}|)$ may be infeasible. Additionally, there are finite sample problems when learning M based on this approach: if an individually fair M is learned from a finite sample of a larger population, M is not guaranteed to remain fair when extended to entire population. These are due in part to the (desirable) strong fairness notion that requires fairness over all individuals. This also motivates the notion of group fairness presented in Section 17.3, which coarsely approximates individual fairness via group membership to achieve better computational efficiency and population-level generalization, at the cost of a less-perfect fairness notion.

The objective $\min \mathbb{E}_{X\sim\mathcal{X}} L(X, \cdot)$ minimizes average loss over the population, but does not optimize for individual loss. The objective could certainly be modified to consider minimax loss over individuals, $\min \max_X L(X, \cdot)$, but this would change

the optimization problem and may harm some of the desirable algorithmic properties described above.

Upshot of Individual Fairness

Perfect in theory; extremely difficult to make practical.

17.2.4 Learning the Metric over Individuals

One of the main barriers to the practical deployment of individual fairness is the requirement of a perfect task-specific metric d over individuals, which is needed to apply the results of Dwork et al. [Dwo+12]. Naturally, access to such a metric is rare to impossible in practice. Instead, Ilvento [Ilv20] proposes approximating a metric for individual fairness using only a small number of queries to a human fairness arbiter, who is "free of explicit biases and possesses sufficient domain knowledge to evaluate similarity," such as a financial regulator or a college admissions officer. This approach builds upon tools from *metric learning*, which aims to automatically construct task-specific distance metrics from weakly supervised data, and also focuses on learning distance metrics from human feedback.

Ilvento [Ilv20] considers two types of queries that can be asked of the human fairness arbiter to learn the distance metric. *Triplet queries* ask whether point x is closer to point y or point z, and *real-valued distance queries* ask for the distance between two points x and y.[2] It is assumed that comparative evaluations are easier for humans than absolute evaluations, and thus distance queries are considered more expensive to ask the arbiter than triplet queries.

A key observation is that individual fairness does not require distances between individuals to be maintained exactly by the Lipschitz mapping, only that the distances are not exceeded—note that the first constraint of (17.1) only requires a one-sided bound that $D(M(x), M(y)) \leq d(x, y)$. This motivates the notion of a *submetric* (Definition 17.6) which is a contraction of the original metric that does not overestimate any distance beyond a small additive error term. Substituting a submetric for the original metric d in [Dwo+12] will still maintain individual fairness and will prove easier to learn.

Definition 17.6 (α-submetric)**.** *Given a metric d, we say that $d' : V \times V \to [0, 1]$ is an α-submetric of d if for all $u, v \in V$, $d'(u, v) \leq d(u, v) + \alpha$.*

2. Relaxations of this framework are also considered, including settings where the arbiter is allowed to respond to triplet queries with "too close to call" if $d(x, y)$ and $d(x, z)$ are very close, and allowing noisy responses to real-valued queries. Similar relaxations were used in Jung et al. [Jun+20], which also considered how to incorporate human expertise into fairness evaluation.

The query-efficient submetric learning algorithm of Ilvento [Ilv20] first randomly selects a small set of representatives from V, then approximately learns the distances between each representative and all other points, and finally combines these to produce a single submetric hypothesis. To learn approximate distances to a single representative, first, points can be sorted in order of increasing distance from the representative using $O(|V| \log |V|)$ inexpensive triplet queries to sort the elements (i.e., "is x or y closer to representative r?"). Given this ordering, $O(\max\{1/\alpha, \log |V|\})$ expensive real-valued distance queries can be used to label the ordering with α-approximate (underestimated) distances.

These approximate distances alone are not sufficient to describe the underlying metric; for example, knowing only that x and y are equidistant from r, it is impossible to distinguish whether $d(x, y)$ is zero or twice their distance from r. Submetrics constructed based on different representatives preserve different information about the underlying metric, so information from all representatives are then aggregated to form a more expressive submetric. Ilvento [Ilv20] provides a method for combining the distances d_r from each representative, and shows that under certain technical assumptions, i.e., how tightly packed individuals are, this approach will result in a submetric with good distance preservation properties.

The approach of Jung et al. [Jun+20], which also focuses on engaging humans in the decision-making process via pairwise comparisons, sidesteps the issue of learning the metric over individuals by not even assuming that such a metric exists. Instead of trying to learn the true underlying metric, they seek to learn a classifier that is consistent with the pairwise fairness constraints elicited from the human arbiter.

Practical applications of these promising theoretical results are still far from immediate. Even though Ilvento [Ilv20] and Jung et al. [Jun+20] show that only a small number of queries are required of the arbiter, any practical system would first have to identify such an unbiased and qualified individual (or collection of individuals) to serve as the human fairness arbiter.

To address the challenge of extending the learned metric to a larger population, Rothblum and Yona [RY18] consider the problem of learning a predictor from a sample of labeled points $(\mathcal{X}, \mathcal{Y})$ that will generalize well to the underlying distribution, rather than learning a classifier for the input set of points. They show that applying the individual fairness definition of Dwork et al. [Dwo+12] as-is for this new goal makes even simple learning tasks infeasible. They instead develop a relaxed approximate metric-fairness notion—formalized as *Probably Approximately Correct and Fair* (PACF), which parallels the definition of PAC learning—that allows for both a small additive slack in the similarity measure of outcomes and a small probability of failure of the fairness guarantee. This allows them to show that (under certain technical conditions) learning an approximately metric-fair predictor

on a sample generalizes to approximate metric-fairness over the underlying distribution.

17.3 Group Fairness

Group fairness focuses on *ensuring fair outcomes with respect to pre-determined protected groups* (e.g., race or gender). Specifically, it assumes that protected groups have been exogenously determined, and divides the attributes of each individual into those which are *protected* (e.g., race, gender) and those which are *unprotected* (e.g., zip code, SAT score). Group fair algorithms are free to use unprotected attributes arbitrarily, but must satisfy (approximately) equal treatment across groups, conditioned on any realization of the protected attribute.

Unlike differential privacy, there is no single unified definition of group fairness; the problem of defining group fairness is itself an active research area, with many natural definitions put forth, corresponding to different mathematical formalizations of "equal treatment across groups" [Nar18; Mit+21]. Unfortunately, several of the most popular definitions have been shown to be incompatible with each other, and it is impossible for any group fairness definition to achieve a short list of natural desiderata [KMR17]. As such, the appropriate formal definition of group fairness should depend on the use case and the analysis task.

The approach of focusing on fairness with respect to protected group membership is in part a response to the practical challenges of individual fairness: although it may be impractical to guarantee fairness across *all* individuals, we will see that standard machine learning tools are well-equipped to construct classifiers that guarantee similar outcomes across a small number of well-defined groups. The focus on group membership is also born out of the reality that in many practical contexts, fair treatment is legally or ethically mandated based on certain protected attributes, such as race, gender, sexual orientation, or disability status.

However, this approach also introduces new practical challenges when protected groups are not automatically defined, since fair treatment is *only* guaranteed with respect to the specified groups. Which groups deserve protection? How should correlations between protected and unprotected attributes be handled? The mathematical formalization of group fairness sidesteps this issue by assuming an exhaustive, pre-defined list of protected groups, which is input to the algorithm. This may be appropriate for applications such as housing or employment, where the set of protected attributes are legally defined, but poses a meaningful barrier for other critical applications, such as loan approval or recidivism prediction, where the division between protected and unprotected attributes is less straightforward.

Relationship to Differential Privacy

Unlike individual fairness, group fairness does not have an inherent technical connection to differential privacy. However, analysts who are concerned with the societal implications of their machine learning pipeline may want to simultaneously achieve privacy and fairness. In this framework, the analyst would require differential privacy for the training dataset that is used to learn the classifier, and fairness for the future test set (possibly the rest of the population) to which the classifier will be applied.

We emphasize that group fairness and differential privacy are orthogonal goals to be satisfied on separate portions of the dataset. The question becomes the compatibility of these two objectives, and whether they can be simultaneously achieved. While the design of algorithms for learning group-fair classifiers has been studied extensively in the fairness literature, we will focus here on the results of two papers [CGKM19; Jag+19] that provide a positive answer to this question by designing *differentially private* algorithms for learning group-fair classifiers. These two papers take a similar algorithmic approach (presented in Section 17.3.2) to achieve privacy for the training set and fairness for the test set.

17.3.1 Setting

Consider the problem of binary classification in a semi-supervised setting. Each individual's data forms a triple (X, A, Y), where X is an unprotected attribute, A is a protected attribute, and Y is a binary label (e.g., will repay a loan or not).[3] These triples are drawn from an unknown joint distribution P, corresponding the underlying population. Given n i.i.d. samples from P, an analyst's goal is to learn a classifier $\hat{Y}$ that maps observable attributes (X, A) to predicted labels[4] $\hat{Y}$ in a way that:

- is *differentially private* with respect to the database of n sampled training points,
- yields a $\hat{Y}$ that is *fair* with respect to the protected attribute A, and
- $\hat{Y}$ is an *accurate* prediction of Y given (X, A).

We emphasize that in contrast to the individual fairness setting, which sought predictions only for the given set of individuals, we move here to a more classical machine learning setting, where we have access to a finite sample from a larger

3. On a (labeled) training dataset, the analyst will be able to observe all (X, A, Y) of an individual, but on the test set, the analyst will only be able to observe (X, A) and must predict $\hat{Y}$.

4. To simplify presentation, we will abuse notation to let $\hat{Y}$ denote both the mapping and the outcome of that mapping on an implied (X, A).

underlying population, and we seek to learn a classifier that can be applied to the entire population.

As discussed above, there are many possible definitions of group fairness. In the setting of binary classification that we consider here, the most commonly used fairness notions are Equalized Odds (Definition 17.8) and Equal Opportunity (Remark 17.9), which require similar false positive and true positive rates across groups [HPS16]. We define these first.

Definition 17.7 (False Positive Rate, True Positive Rate)**.** *Let $FP_a(\hat{Y})$ and $TP_a(\hat{Y})$ respectively denote the false positive rate and true positive rate of classifier $\hat{Y}$ on the group $\{A = a\}$:*

$$FP_a(\hat{Y}) = \mathbf{Pr}[\hat{Y} = 1 | A = a, Y = 0]$$
$$TP_a(\hat{Y}) = \mathbf{Pr}[\hat{Y} = 1 | A = a, Y = 1],$$

where the probabilities are taken over the distribution P and the randomness of the classifier.

We also define empirical false positive rate and true positive rate, $\widehat{\mathrm{FP}}_a(\hat{Y})$ and $\widehat{\mathrm{FP}}_a(\hat{Y})$ as the average rates evaluated on the labeled sample.

We use *Equalized Odds* as our fairness notion, which requires all subgroups $a \in A$ to have similar true and false positive rates. Similar *true* (resp., *false*) positive rates implies that conditioned on being a qualified candidate with $Y = 1$ (resp., unqualified candidate with $Y = 0$), the chance of getting the good outcome $\hat{Y} = 1$ is approximately equalized across groups.

Definition 17.8 (γ-Equalized Odds fairness)**.** *A classifier $\hat{Y}$ satisfies γ-equalized odds fairness with respect to protected attribute A if $\forall a, a' \in A$, both the following conditions hold:*

$$|FP_a(\hat{Y}) - FP_{a'}(\hat{Y})| \leq \gamma \quad \text{and} \quad |TP_a(\hat{Y}) - TP_{a'}(\hat{Y})| \leq \gamma\,.$$

Remark 17.9. *A slightly relaxed fairness notion is that of* Equal Opportunity*, which only requires similar true positive rates across groups. This relaxation only requires that qualified candidates ($Y = 1$) receive fair treatment, and makes no fairness requirement for the unqualified candidates ($Y = 0$).*

To measure accuracy, define the *error* of a classifier $\hat{Y}$ to be its misclassification error on the population:

$$\mathrm{err}(\hat{Y}) = \mathop{\mathbf{Pr}}_{P,\hat{Y}}[\hat{Y} \neq Y].$$

We also define empirical error $\widehat{\text{err}}(\hat{Y})$ as the average misclassification error on the training set. Given a hypothesis class $\mathcal{H}$ (e.g., the set of all binary classifiers), we will consider randomized classifiers $\hat{Y}$ in $\Delta(\mathcal{H})$.

The analyst's formal goal is to find a classifier $\hat{Y}$ that minimizes empirical error subject to the equalized odds fairness constraint, which can be formalized as the following optimization problem:

$$\begin{aligned} &\min_{\hat{Y} \in \Delta(\mathcal{H})} \widehat{\text{err}}(\hat{Y}) \\ \text{s.t. } &|\widehat{\text{FP}}_a(\hat{Y}) - \widehat{\text{FP}}_{a'}(\hat{Y})| \leq \gamma \quad \forall a, a' \in A \\ &|\widehat{\text{TP}}_a(\hat{Y}) - \widehat{\text{TP}}_{a'}(\hat{Y})| \leq \gamma \quad \forall a, a' \in A. \end{aligned} \tag{17.2}$$

Solving the optimization problem of (17.2) will ensure a classifier that is accurate and satisfies γ -Equalized Odds group fairness. Of course, the analyst would prefer to minimize distributional error subject to distributional true and false positive rates, but they must instead settle for the empirical proxies evaluated on their finite sample. This gap will be accounted for in the final approximation bound given in Section 17.3.2.

If we additionally use a differentially private method for solving the optimization problem, then we will satisfy the desiderata of privacy and fairness. Fortunately, the requisite private machine learning and optimization tools exist, which we will see next.

17.3.2 Algorithmic Approach and Results

Since the optimization problem of (17.2) encodes the goal of maximizing accuracy subject to a group fairness constraint, the remaining task is only to solve (17.2) in a differentially private and computationally efficient way. Both Cummings et al. [CGKM19] and Jagielski et al. [Jag+19] take the same high-level approach to this task, with the following three steps: (1) re-write (17.2) as a linear program of finite size; (2) formulate the LP as a two-player zero-sum game; and (3) solve the game with differentially private no-regret learning dynamics.

This approach relies on a foundational result of Freund and Schapire [FS96], which states that in two-player zero-sum games, if one player plays according to a no-regret learning algorithm and the other player best-responds, then average play of both players will converge to a Nash equilibrium of the game. By strong duality, this equilibrium corresponds to an optimal solution of the LP, which must be a fair and accurate classifier. Since the learning algorithm is differentially private, this also satisfies the privacy requirement.

Below we present an informal statement of this main result, and then proceed with details of the algorithmic construction.

Theorem 17.10 (Cummings et al. [CGKM19] and Jagielski et al. [Jag+19])**.** *There exists an* (ε, δ)*-differentially private algorithm that runs in polynomial time (given access to an oracle), and with probability at least* $1 - \beta$ *outputs a random classifier* $\hat{Y}$ *that satisfies* $err(\hat{Y}) \leq OPT + \alpha$ *and* $(\gamma + \alpha)$*-Equalized Odds, for* $\alpha = O(\sqrt{\frac{\log(1/\beta)}{n\varepsilon}})$.

Step 1: Re-writing the Problem

First we observe that the optimization problem in (17.2) is a linear program with polynomially many constraints. Specifically, note that when a classifier $\hat{Y}$ is described explicitly by the probability of producing $\hat{Y} = 1$ given observable attributes $(X = x, A = a)$, then $\widehat{err}(\hat{Y})$, $\widehat{\mathrm{FP}}_a(\hat{Y})$, and $\widehat{\mathrm{TP}}_a(\hat{Y})$ can all be expressed as a linear function of the classifier.

For a given $\hat{Y}$, we can write the fairness constraints in matrix form, as $\vec{r}(\hat{Y}) \leq \vec{0}$, where,[5]

$$\vec{r}(\hat{Y}) = \begin{bmatrix} \widehat{\mathrm{FP}}_a(\hat{Y}) - \widehat{\mathrm{FP}}_{a'}(\hat{Y}) - \gamma \\ \widehat{\mathrm{TP}}_a(\hat{Y}) - \widehat{\mathrm{TP}}_{a'}(\hat{Y}) - \gamma \end{bmatrix}_{a,a' \in A} \in \mathbb{R}^{2|A|^2}.$$

Next we can Lagrangify these constraints by introducing a Lagrangian variable for each constraint:

$$\vec{\lambda} = \begin{bmatrix} \lambda_{(FP,a,a')} \\ \lambda_{(TP,a,a')} \end{bmatrix}_{a,a' \in A} \in \mathbb{R}^{2|A|^2}.$$

We assume $\|\vec{\lambda}\|_1 \leq B$ for a given B to ensure convergence of the no-regret dynamics.

Finally, our fair learning LP can equivalently be written as the following Lagrangian minimax problem:

$$\min_{\hat{Y} \in \Delta\mathcal{H}} \max_{\|\vec{\lambda}\|_1 \leq B} \widehat{\mathrm{err}}(\hat{Y}) + \vec{\lambda}^\top \vec{r}(\hat{Y}). \tag{17.3}$$

Step 2: Formulating as a Game

Kearns et al. [KNRW18] was the first to give a reduction from the problem of learning a fair classifier as formalized in (17.3) to the problem of computing an

5. $|A|$ is commonly assumed to be a small constant corresponding to the number of demographic groups in a population. However, if $|A|$ is infinite or otherwise too large for $2|A|^2$ constraints to be feasible, an alternative formulation relying on Sauer's Lemma requires only $O(n^{VC-DIM(A)})$ constraints.

approximate equilibrium of a two-player zero-sum game. This approach uses Sion's Minimax Theorem to write (17.3) as:

$$\min_{\hat{Y}\in\Delta\mathcal{H}} \max_{\|\vec{\lambda}\|_1\le B} \widehat{\mathrm{err}}(\hat{Y}) + \vec{\lambda}^\top \vec{r}(\hat{Y}) = \max_{\|\vec{\lambda}\|_1\le B} \min_{\hat{Y}\in\Delta\mathcal{H}} \widehat{\mathrm{err}}(\hat{Y}) + \vec{\lambda}^\top \vec{r}(\hat{Y}) = OPT, \tag{17.4}$$

where OPT is the optimal objective value to the fair learning problem.

This provides the payoff matrix for a two-player zero-sum game. The primal (minimization) player is a Learner who chooses a classifier $\hat{Y} \in \mathcal{H}$ to minimize empirical error (plus a fairness penalty term given by the Lagrangian). The dual (maximization) player is an Auditor who chooses a fairness constraint that is maximally violated. Intuitively, the Auditor is trying to identify a group $a \in A$ that experiences the largest fairness violation under the Learner's $\hat{Y}$, and puts all the weight on the dual variable corresponding to that constraint. Allowing both players randomized strategies in this game yields the full decision space for the optimization problem.

When these two players iteratively and repeatedly play this zero-sum game, and one plays according to a no-regret learning algorithm while the other plays a best-response, then the foundational result of Freund and Schapire [FS96] guarantees that average game play will converge to an approximate equilibrium. By (17.4), an equilibrium of this game corresponds to an optimal solution of the LP—i.e., a fair and accurate classifier.

Step 3: Private No-regret Dynamics

While the roles of "no-regret algorithm" and "best-responder" are interchangeable between players,[6] both players must compute their action differentially privately. The Learner relies on the training data to evaluate classifier accuracy, and the Auditor relies on the training data to find a fairness constraint that is violated.

The player using the no-regret algorithm can use any of the numerous differentially private no-regret learning algorithms—such as Private Follow the Perturbed Leader or Private Multiplicative Weights—to compute their action. For the other player, Kearns et al. [KNRW18] show in the non-private setting that the best-response of both players can be reduced to cost-sensitive classification (CSC), where the cost of predicting a label depends on the label value. Thus the algorithmic results rely on the existence of a differentially private CSC oracle—such as the Exponential Mechanism or another private heuristic—to allow the best-response player to efficiently and privately compute their action.

6. Cummings et al. [CGKM19] has the Learner play a no-regret algorithm and the Auditor best-responds, while Jagielski et al. [Jag+19] uses the reverse roles.

Thus we have privately learned a fair and accurate classifier, as desired. The additional approximation terms in the accuracy and fairness guarantees in Theorem 17.10 come from inherent noise in the private no-regret learner and the private CSC oracle. Additionally, Freund and Schapire [FS96] only guarantee an *approximate*-equilibrium, which corresponds to an approximate solution to (17.2). Finally, since (17.2) necessarily evaluates accuracy and fairness empirically based on the training data, there will be some loss that depends on the sample size when generalizing back to the original (infinite) population.

Remark 17.11. *Alabi [Ala19] provides improved sample complexity bounds for the problem of private and group fair learning, under the assumption of a slightly different oracle. Those bounds apply to Equalized Odds fairness, as well as other group fairness notions that can be expressed as a convex loss function.*

17.3.3 Pros and Cons of Group Fairness

Pros

The most obvious advantage of focusing on group fairness is that it integrates will with existing machine learning frameworks, and can be achieved in a computationally efficient way using existing algorithmic tools. This lowers the barriers for use, and makes it easier for theorists to design group fair algorithms, and for practitioners to implement them.

While this chapter focused primarily on Equalized Odds as a notion of group fairness, it is far from the only definition; dozens have been proposed, discussed, and used in the literature (see, e.g., Narayanan [Nar18] and Mitchell et al. [Mit+21] for surveys). Group fairness encompasses a family of fairness definitions, all sharing the same general flavor: "the treatment of individuals should independent of their protected attributes." This family ranges from attribute-blind policies (e.g., college admission based only on SAT scores), to attribute-aware policies such as affirmative action to make up for historical inequities, to dynamic policies that account for the many decisions made about an individual over the course of her life. The specific definition of group fairness used in any particular application can be tailored to suit practical needs, based on domain-specific fairness concerns and modeling choices.

Cons

The myriad of definitions is also a weakness of group fairness. There is no one single correct definition of group fairness that theorists or practitioners can use off-the-shelf across all application domains. Instead, they must understand the contextual and sociological requirements that give rise to fairness concerns in their specific application domain, and choose the appropriate mathematical fairness definition. This challenge is exacerbated by the fact that many common group fairness notions

are incompatible, and provably cannot be achieved simultaneously [KMR17]. Thus while the field has readily available group-fair algorithms, the deployment of these tools does not scale easily to new use cases.

Relatedly, group fairness requires the analyst to pre-specify all groups requiring fair treatment as an input to the algorithm, and it does not provide any fairness guarantees with respect to other groups. In the absence of laws that explicitly enumerate legally protected attributes, determining the appropriate collection of protected groups is far from straightforward. This challenge is exacerbated when seemingly unimportant attributes may serve as a proxy for the protected attribute (e.g., shampoo choice may be highly correlated with race and gender). With high-dimensional training datasets and complex machine learning algorithms, it is easy to inadvertently encode bias through proxies. The selection of protected groups—like the choice of fairness definition—requires significant sociological research and domain expertise, which again slows the deployment of group-fair algorithms.

Finally, group fairness ensures fairness *on average* for a group, but does not provide guarantees across individuals within a group. Within a fixed group (corresponding to some $a \in A$), more qualified individuals may receive worse outcomes than less qualified individuals, even under a group-fair classifier. Within-group unfairness enables other types of unfair treatment such as reverse tokenism, where the least qualified members of a group are selected as positive examples (i.e., $\hat{Y} = 1$) to spuriously demonstrate lack of merit for the entire group, and it ignores intersectionality, where individuals may belong to multiple overlapping groups (e.g., based on race, gender, and disability status). Satisfying group fairness with respect to each group independently is not sufficient to guarantee fairness across the intersection of these groups, as we demonstrate next in Section 17.4.

Upshot of Group Fairness

Easy to achieve algorithmically; doesn't capture all fairness concerns.

17.4 Multi-group Fairness

Two of the shortcomings of group fairness as studied in Section 17.3 are that it requires all protected groups to be specified in advance, and that it does not capture notions of *intersectionality*, where individuals may belong to multiple overlapping groups—for example, a group corresponding to their race and a group corresponding to their gender. In the computer science literature, the latter concern is referred to as *multi-group fairness*.

As an illustration of the need for multi-group fairness, consider the following example. Imagine all people have two protected attributes: shape—round or

square—and color— blue or green.[7] Imagine that these attributes are uniformly distributed in the population and independent of an individual's ability or intent to repay a loan. Consider the classifier that awards loans to everyone who is round-blue and square-green, and denies all others. This classifier would satisfy perfect Equalized Odds fairness (Definition 17.8) with respect to shape and color separately, but not when the two features are considered together. In particular, this would be unfair to round-green and square-blue people under any reasonable definition of fairness, since they have no hope of a positive outcome, regardless of their merit. Similar examples were given in [DI19], which showed that group-fair classifiers do not *compose* in the way that differential privacy does. For example, making fair hiring decisions with respect to race and gender independently does not automatically ensure that combinations thereof will be treated fairly.

One solution for this simple example is to redefine protected attributes to include these intersections, and require group fairness across all four newly-defined groups. However, the computational complexity of this approach scales exponentially with the number of protected attributes, which may not be feasible in many practical applications, particularly when protected attributes take on non-binary (e.g., income) or continuous values (e.g., probability of developing a disease). Additionally, this would maintain the weakness of group fairness that one must pre-specify all possible groups that require fair treatment; in practice, the full set of protected attributes may not be known in advance, particularly due to correlations across attributes. E.g., hair color is not a legally protected attribute, but it can be a proxy for race due to correlations between race and hair color.

In this section, we will see approaches for multi-group fairness that do not require the set of protected groups to be known in advance, and are computationally efficient, even for an arbitrarily large number of protected groups.

17.4.1 Setting

In the multi-group fairness setting, each individual's data consists of a pair (X, Y) drawn from an unknown distribution P, where X includes all attributes, and Y is a binary label. Note that we no longer specify protected attributes A, but instead introduce an arbitrary attribute domain $\mathcal{X}$ that allows for high-dimensional attribute vector X. It will be the implicit task of the learning algorithm to determine which combinations of attributes require protection.

Our learning goal is a continuous *prediction task*—a relaxation of the binary classification task considered in Section 17.3. Given n i.i.d. samples from P, the

7. For a more provocative example, consider these attributes to be race and gender. To avoid addressing complex social issues which are beyond the scope of this book, we use made-up attributes instead.

analyst would like to learn a predictor $p : \mathcal{X} \to [0, 1]$, where the prediction $p(X) \in [0, 1]$ can be interpreted as the predicted probability of the event that $Y = 1$ for an individual with attributes X.

Note that the Bayes optimal predictor $p^*(X) = \mathbf{Pr}[Y = 1|X]$ both maximizes predictive accuracy, and is perfectly fair to all individuals. That is, it is individually fair with respect to the metric over individuals $d(X_i, X_j) = |\mathbf{Pr}[Y = 1|X_i] - \mathbf{Pr}[Y = 1|X_j]|$. However, learning this optimal predictor is information theoretically impossible from a finite sample. Instead, we will seek to learn a predictor that seeks to maximize both predictive accuracy and the number of groups that receive fairness protections, given information-theoretic and computational constraints.

As we will see, the method for non-privately learning a multi-group fair predictor has natural algorithmic parallels to learning a group-fair classifier in Section 17.3, while known private methods for this problem take a substantially different algorithmic form. As a result, we will primarily focus here on algorithms for learning a multi-group fair predictor without a privacy constraint in order to better highlight this connection, with a discussion of differentially private techniques for learning a multi-group fair predictor deferred to Section 17.4.4.

17.4.2 Multi-calibration

A common fairness solution for predictors is *calibration*, which attempts to ensure that groups of individuals with similar risk, $\mathbf{Pr}[Y = 1|X]$, receive similar predictions, $p(X)$. Calibration is also commonly used (outside of fairness) as an accuracy notion in the statistics and forecasting literature (e.g., [SSV03]) to ensure that the probabilistic prediction of an event occurring is close to the true probability.

Definition 17.12 (α-calibration)**.** *A predictor $p : \mathcal{X} \to [0, 1]$ is* α-calibrated *if for all $v \in supp(p)$,*

$$|\mathbf{Pr}[Y = 1|p(X) = v] - v| \leq \alpha,$$

or equivalently, in terms of the Bayes optimal predictor $p^*(X)$,

$$|\mathbb{E}[p^*(X)|p(X) = v] - v| \leq \alpha.$$

In words, calibration requires that of all the individuals who receive prediction v, their average positive outcome probability is close to v—e.g., among the people who receive a prediction of 70% chance of having $Y = 1$, 70% of them should truly have $Y = 1$. Note that for Boolean predictors, exact calibration ($\alpha = 0$) requires perfect accuracy.

However, while calibration is a desirable property, it is a relatively weak condition, and it alone is not sufficient to ensure fairness. Concretely, the *mean-predictor* $p_\mu(X) = \mu := \mathbb{E}_P[Y]$ is perfectly calibrated but uninformative: $\mathbf{Pr}[Y = 1 | p_\mu = \mu] = \mathbf{Pr}[Y = 1] = \mu$. To relate this to fairness, consider two groups S and T with identical distributions of Y. Imagine that the predictor used for group S is calibrated and informative, where some individuals receive predictions above the mean, and that group T simply receives the (perfectly calibrated) mean predictor. Any reasonable decision-making procedure based on these predictions, will provide different outcomes to individuals with the same risk across different groups, due to their differently ranked predictions. This problem will be particularly exacerbated for minority groups who are underrepresented in training data, for whom an informative predictor will be more difficult to learn.

As with the binary classification setting studied in Sections 17.2 and 17.3, we observe that calibration as a fairness notion has a gap in treatment between individual and group fairness. Calibration across a small number of pre-defined groups is easily achievable, but does not provide sufficient fairness, as too many diverse individuals may be grouped together under one single prediction value. At the other extreme, individual calibration provides the guarantee that the prediction made for each individual is her true probability of receiving a 1-label: $\mathbf{Pr}[Y = 1 | p(X) = v \wedge X] = p^*(X)$. However, this requires learning the Bayes optimal predictor, which is unattainable from a finite sample.

This motivates the definition of *multi-calibration*, which bridges between these two notions to provide multi-group fairness. Multi-calibration will guarantee calibration across a set of computationally identifiable groups, where the notion of *identifiability* will depend on a parameterized computational bound. This provides a computationally efficient relaxation of individual fairness, and it also strengthens the notion of group fairness by going beyond a small number of pre-defined protected groups.

Definition 17.13 (Multi-calibration [HKRR18])**.** *Let* $\mathcal{C} \subseteq \{c : \mathcal{X} \to \{0, 1\}\}$. *A predictor* p *is* $(\mathcal{C}, \alpha)$-multi-calibrated *if* $\forall c \in \mathcal{C}$ *and* $\forall v \in supp(p)$,

$$|\mathbb{E}[Y | p(X) = v \wedge c(X) = 1] - v| \leq \alpha.$$

The class $\mathcal{C}$ describes the collection of subpopulations that require privacy protections, where for each subpopulation $\mathcal{S} \subseteq \mathcal{X}$, there should exist a function $c_\mathcal{S}$ such that $c_\mathcal{S}(X) = 1$ if and only if $X \in \mathcal{S}$. The class $\mathcal{C}$ should be chosen to be as expressive as the analyst can afford computationally—e.g., the class of functions that are implementable by depth-d decision trees. This gives the analyst explicit control over the balance between group and individual fairness—a larger and richer

$\mathcal{C}$ will result in fairness across more protected groups, but will be more computationally expensive.

An alternative interpretation of the role of $\mathcal{C}$ is that in order to claim that a predictor is unfair, multi-calibration requires a "witness" of a group that is treated unfairly. If $\mathcal{C}$ is large, this grants more computational power to find such a witness, which also requires fairness across more groups.

One benefit of multi-calibration is that it ensures that no qualified subpopulations in $\mathcal{C}$ are overlooked. For example, suppose there exists $c \in \mathcal{C}$ such that $\mathbb{E}[Y|c(X) = 1] > 1 - \alpha$. Then any $(\mathcal{C}, \alpha)$-multi-calibrated predictor should give high predictions on the group described by c. Thus $\mathcal{C}$ identifies the prediction-relevant structure in $Y|X$. This also ensures that multi-calibration is robust to under-representation, as predictions must be calibrated even if a group is small. Additionally, it requires learning *within* protected groups to identify qualified individuals a priori, thus avoiding the pitfalls illustrated in the example above with groups S and T.

17.4.3 Learning Multi-calibrated Predictors

The algorithmic process for learning a multi-calibrated predictor involves a similar primal-dual framework as the algorithm in Section 17.3.2 for learning a group-fair classifier, with some key technical differences. The high-level process is still: (1) write the problem formally as a constrained optimization problem; (2) formulate the optimization problem as a two-player zero-sum game; and (3) solve the game using no-regret learning dynamics.

As in Section 17.3.2, this approach will rely on the result of Freund and Schapire [FS96], that average no-regret play of a two-player zero-sum game converges to an approximate Nash equilibrium. With the appropriate game formulation in Step 2, an equilibrium strategy of the Learner (primal player) will correspond to a multi-calibrated predictor.

The main difference from Section 17.3.2 is that the underlying optimization cannot be written as an LP, but instead will be a non-linear feasibility problem. This new optimization problem leads to a modified no-regret learning set-up, and in particular, requires a reduction to weak agnostic learning for the Auditor to compute a best-response.

Nevertheless, the following theorem says that a multi-calibrated predictor can be efficiently learned using this procedure. The second part of the result tells us that the learned predictor can be described with a circuit of size not much larger than the complexity of representing the functions in $\mathcal{C}$.

Theorem 17.14 ([HKRR18])**.** *There exists an efficient algorithm that, if given access to a Weak Agnostic Learner, learns a* $(\mathcal{C}, \alpha)$*-multi-calibrated predictor* p *using*

$O(\log|\mathcal{C}|\text{poly}(1/\alpha))$ *labeled samples in* $O(|\mathcal{C}|\text{poly}(1/\alpha))$ *time. Further, if membership in each set* $c \in \mathcal{C}$ *can be evaluated by a circuit of size* s*, then* p *can be implemented by a circuit of size* $O(s \cdot \text{poly}(1/\alpha))$.

This result also says that learning a multi-calibrated predictor can be interpreted as a boosting algorithm, as weak agnostic learners can be boosted to strong agnostic learners after polynomially many iterations. The Auditor is a weak learner that identifies sources of unfairness in the Learner's current predictor. If the Auditor identifies a group that is treated unfairly, then this is used as an update that makes significant process towards multi-calibration. If the Auditor fails to identify such a group, then the current predictor must be multi-calibrated.

Now we proceed with details of the algorithmic construction.

Step 1: Framing the Optimization Problem

The major difference from Section 17.3.2 is that there is no explicit error objective in the optimization problem to be solved.[8] Under group fairness, the goal was to maximize accuracy (that is, minimize error) subject to a fairness constraint. Here, multi-calibration is an accuracy notion—ensuring that average predictions on subgroups are correct—so we do not need to separately track accuracy and fairness. This leaves us with the following feasibility problem, where we need only to find a predictor that satisfies the multi-calibration constraint:

$$\min_{p:\mathcal{X}\to[0,1]} 0 \tag{17.5}$$
$$\text{s.t. } |\mathbb{E}[Y|p(X) = v \wedge c(X) = 1] - v| \leq \alpha \quad \forall c \in \mathcal{C} \text{ and } \forall v \in \mathit{supp}(p).$$

Step 2: Formulation as a Game

While we can no longer explicitly Lagrangify the constraint in this formulation due to the the requirement that the prediction error must be below α for all $v \in \mathit{supp}(p)$, we can still view this as a two-player zero-sum game between a Learner and an Auditor.

The primal player is a Learner who chooses a predictor p, and the dual player is an Auditor who attempts to find a sub-population identified by $c \in \mathcal{C}$ that is treated unfairly under p. The payoff to the Auditor is the calibration error α on the sub-population c under the predictor p; the Learner's payoff is $-\alpha$.

Step 3: No-regret Dynamics

The Learner is facing a traditional online learning set-up when this game is played repeatedly: they can use any no-regret learning algorithm, iteratively update the

8. This objective is not needed because the goals of multi-calibration are already aligned with regression accuracy, and imply standard loss minimization [Gop+22].

chosen predictor based on the bandit feedback from the Auditor, and (under the assumption of accurate feedback from the Auditor) this approach will converge to a solution where the Auditor cannot identify any (computationally-efficiently identifiable) violated sub-populations, which implies a multi-calibrated predictor. All that remains to be shown is that such an algorithm exists for the auditor.

Formally, the Auditor faces the following problem: Given $p : \mathcal{X} \to [0, 1]$, find a $c \in \mathcal{C}$ such that $|\mathbb{E}[Y|p(X) = v \wedge c(X) = 1] - v|$ is large, or equivalently, such that $|\mathbb{E}[c(X)(Y - v)]|$ is large. Note that this is exactly the problem of Weak Agnostic Learning for correlation detection, where a learner wants to identify correlations greater than α between X and $Y - p(X)$. This can be formalized through the following equivalence of auditing a multi-calibrated learner and Weak Agnostic Learning.

Theorem 17.15 ([HKRR18])**.** *If $\mathcal{C}$ is α-weakly agnostically learnable, then there exists a $(\mathcal{C}, \alpha)$-multi-calibrated learner. Similarly, if there exists a $(\mathcal{C}, \alpha)$-multi-calibrated learner, then $\mathcal{C}$ is α-agnostically learnable.*

Thus auditing multi-calibrated predictors is identical to weak agnostic learning (i.e., correlation detection), and by a reduction, we can go from auditing (or weak agnostic learning) to learning a multi-calibrated predictor. This gives us an algorithm to satisfy the learning goal of (computationally efficiently) identifying any function that satisfies fairness in the form of calibration for the sub-populations of interest, as specified by the class $\mathcal{C}$.

More concretely, in the repeated game, the Learner (primal player) can use any no-regret learning algorithm to learn a multi-calibrated predictor, and the Auditor (dual player) uses a Weak Agnostic Learning oracle to identify a sub-population in each round. No-regret dynamics and the result of Freund and Schapire [FS96] ensure that average play converges to an approximate-Nash equilibrium. The average predictor (i.e., averaged over all rounds of play) of the Learner corresponds to an approximately optimal solution to the original optimization problem (17.5), which will also be a multi-calibrated predictor.

Remark 17.16. *From a practical perspective, we note that while agnostic learning is computationally hard (i.e., it is equivalent to boosting), it can be efficient in some cases. For example, when predictions are in* $[0, 1]$ *rather than Boolean, this is a regression problem, which can be solved efficiently. More broadly, nearly all of machine learning is based on agnostic learning, so relying on a weak agnostic learner should work well in practice, despite theoretical hardness.*

Remark 17.17 (Individual calibration)**.** *The framework in this section considers settings where each individual participates in this experience only once (e.g., applying for a*

mortgage), and experiences a single (high-stakes) binary outcome. Kearns et al. [KRS19] consider a variant setting where each individual instead participates in many low-stakes labelings (e.g., impressions of targeted ads). In this case, considering individual fairness in expectation may be justified since the individual can realize the outcomes of many personalized decisions.

This set-up can be viewed equivalently as an average-case version of individual fairness as formalized in Section 17.2, or as an individual version of calibration as formalized here, where each individual's distribution of outcomes should be calibrated with respect to their personal ground truth.

17.4.4 Relationship to Differential Privacy

The algorithmic approach presented in Section 17.4.3 for learning multi-calibrated predictors is not differentially private as stated, although there are many indications that it likely could be made differentially private, to enable privately-trained multi-group fair predictors. Firstly, the original algorithmic construction for learning a multi-calibrated predictor "borrows ideas from the literature on differentially private query release and optimization" [HKRR18]. This suggests an underlying relationship between multi-calibration and differential privacy, even though the algorithm itself is not fully differentially private due to the use of other non-private subroutines. Secondly, the overall framework for learning a multi-calibrated predictor in Section 17.4.3 closely parallels the framework for learning a group-fair classifier in Section 17.3.2. Both involve first framing the fairness-constrained learning problem as a minimax optimization problem, then formulating the optimization problem as a two-player zero-sum game that can be solved by a no-regret learner, and finally showing that an equilibrium of the game corresponds to a solution to the original learning problem. In Section 17.3.2, this process is made private through the use of a private no-regret learner, which suggests that a private no-regret learner could similarly be used to privately learn a multi-calibrated predictor.

This open problem was answered in the affirmative by Kim [Kim20], which showed that multi-calibrated predictors can be learned in a differentially private way. Rather than adding differential privacy using the method above, Kim [Kim20] took an alternative approach that results in improved sample complexity of learning in some parameter regimes. This approach relies on statistical query (SQ) learning, and in particular, differentially private SQ oracles and differentially private density estimation oracles. The algorithm iteratively cycles through all classes in $\mathcal{C}$ and all possible outputs v, and (privately) checks whether the density at v is sufficiently high to be worth investigating the predictor's correctness at v. If yes, it calls a (private) SQ oracle to check whether the oracle gives an answer that is sufficiently

far from the current prediction. If so, then this (private) SQ oracle is used to update the predictor; if not, then the algorithm continues checking classes and outputs.

This algorithmic approach is similar in structure to Private Multiplicative Weights (PMW) (see Section 4.6.1 of Chapter 4 for details), and its privacy analysis also follows a similar structure: it is shown that not too many update steps are required for convergence to a good predictor, and that each update step is differentially private; thus the the overall privacy guarantee follows from composition. Because the algorithmic approach is sufficiently different to other approaches presented in this chapter, and the analysis is quite subtle, it is not presented here; the interested reader is instead referred to [Kim20].

17.5 Impact of Privacy on Fair Outcomes

In an ideal world, privacy and fairness desiderata would be jointly considered to create algorithms that are private and fair by design. This is what we have seen in the previous sections, which focused on privately learning fair classifiers and predictors by considering these two objectives together. However, when privacy and fairness are considered separately—or, more generally, when they are considered independently of the complex system where they are used—the desired outcome may not be achieved. In this section, we will see several stylized examples of real-world applications where this occurs, as well as insights as to the causes and potential methods for remediation. While this is an active field of research that is rapidly evolving, we will survey what is currently known at time of writing as well as open research directions in the field.

Section 17.5.1 considers the impact of privacy noise on fairness of accuracy across groups, and shows that smaller groups tend to receive lower accuracy under a classifier learned via DP-SGD. This unfairness is due in part to fundamental limits of differential privacy which, by definition, limits the impact of small groups on learning, as well as due to specific details of the DP-SGD algorithm. Section 17.5.2 shows that adding differential privacy may cause unfairness in downstream decision-making tasks such as resource allocation, threshold comparison, and regression, all inspired by the U.S. Census Bureau's use of differential privacy. For each decision-making task, we see the cause of unfairness in terms of the privacy noise, and discuss potential solutions to address these disparities. Finally, Section 17.5.3 focuses on *fairness composition* as an analogy to privacy composition, and shows that unlike privacy, fairness does not compose. Several real-world examples are presented where fair classifiers do not compose into fair systems.

17.5.1 Differential Privacy's Impact on Model Accuracy Across Groups

When private learning tools are applied in practice without explicit fairness corrections, it is possible that differential privacy may have a disparate impact across subpopulations, with lower accuracy on smaller groups. [BPS19] demonstrated this effect empirically for deep learning models trained via a differentially private version of stochastic gradient descent (DP-SGD), where privacy of the gradient update step is achieved by first *clipping* the gradient to have a bounded norm, and then adding Gaussian noise that scales with $1/\varepsilon$ and the clipping parameter. See Section 6.3.1 for more details of the algorithm and its privacy guarantees.

[BPS19] found that accuracy of DP-SGD-trained models was not equal across all groups, but rather underrepresented classes in the training data received lower classification accuracy from the resulting model. They also found that while these accuracy disparities exist for non-private models, the extent of the disparity was worse under a privately trained classifier relative to a non-privately trained one. For example, an age and gender classification model trained using DP-SGD ($\varepsilon = 5.69$) had much lower accuracy for faces with darker skin tones (∼59%) than for lighter skin tones (∼78%), whereas the comparable non-private model achieved ∼90% accuracy on lighter skin tones vs. ∼85% on darker skin tones. Similar disparities and differences between private and non-private models persisted across other datasets and other classification tasks, suggesting that adding DP exacerbates existing unfairness in machine learning pipelines.

From a philosophical privacy perspective, this finding should be unsurprising. The goal of differential privacy is to hide the effect of one sample point; for populations represented by fewer points in the training data, the effect of this population on the model is more likely to be obfuscated by the group privacy properties of differential privacy, and thus private models are less likely to capture relevant features for accurately classifying statistical-minority[9] groups. Even without privacy, one would expect machine learning models to have poorer performance on subgroups that are not sufficiently represented in the training data, because including additional training points generally lead to better performance of the learned model. The interesting observation of [BPS19] is that the additional disparity from privacy appears to be substantially larger than the disparity from just differences in sample sizes alone.

9. It is important to distinguish between *minority groups* in society, corresponding to historically marginalized groups or protected social groups, and *statistical minorities*, corresponding to groups that are underrepresented in the dataset, which does not have any normative or social overtones.

To better understand how widely these phenomena will occur in general learning settings, future work is needed to disentangle how much of the observed effect is due to the algorithmic specifics of DP-SGD, versus fundamental limitations of the differential privacy constraint for accurate learning on minority groups. [BPS19] show that certain details the DP-SGD algorithm are one root cause of this disproportionate impact on model accuracy for underrepresented classes. First, points from underrepresented classes are simply less likely to be sampled in the gradient update step under uniform sampling. Even when they are sampled, these are precisely the points that will have a large gradient update and will be clipped, thus removing relevant update information. Without noise, these gradient updates would eventually converge to an accurate model, although perhaps at a slower rate. However, with both clipping and noise addition, the clipped gradients from the underrepresented class are not large enough to compensate for the noise and to allow the model to meaningfully update on this under-represented population. The highlighted text should be replaced with: [TDF21] also show that in DP-SGD, clipping and noise addition affect the gradient norms of different groups differently, which causes unfairness. It remains an open question to what extent the disparate performance across groups— beyond that which exists with non-private learning—is present and unavoidable under other private learning algorithms as well.

17.5.2 Differential Privacy's Impact on Downstream Fair Decision-making

The performance of a differentially private algorithm is traditionally measured in terms of additive accuracy between its output, such as distributional parameters estimated on the training dataset, and some ground truth, such as the empirical parameter value on the training data or the true value on the underlying distribution. However, the output of DP algorithms may be used for downstream decision tasks, where the domain-appropriate fairness notion may have a non-linear relationship with additive accuracy.

One notable and relevant use-case is the 2020 Decennial Census, where differentially private analyses of the collected data are used to allocate a fixed pool of federal resources. The noise added to preserve privacy inherently introduces small inaccuracies in the results; the related fairness question is whether some groups disproportionately bear the cost of these inaccuracies. [Puj+20] and [CDMS21] studied the downstream effects of the inaccuracies introduced by privacy noise in the Census use-case, with a particular emphasis on fairness in outcomes of decision-making problems.

[Puj+20] focused on three real-world *assignment problems* based on Census data products: (1) allocating funds to school districts, (2) voting rights benefits (i.e.,

translations of voting materials for minority language groups), and (3) apportionment of Congressional seats. The authors used public-use Census data as the ground truth;[10] they solved each assignment problem both using the non-private data and using differently private statistics computed on the dataset, and compared the outcomes using problem-specific fairness notions. They found that privacy noise can have a large impact, particularly when there is a misalignment of the upstream accuracy guarantees of the DP algorithms with the downstream fairness metric.

For educational funding, fairness was measured as a correct fractional allocation of funds relative to the ground truth. With privacy, smaller districts tend to get an inflated allocation of funds, at the expense of larger districts losing a small portion of their deserved funds. The effects were particularly noticeable for the smaller school districts, which may be due in part to the mismatch between the additive accuracy guarantees of most DP algorithms with the multiplicative fairness metric, which will cause seemly larger effects in smaller populations. Voting rights benefits is the problem of classifying minority language group populations as above or below the size threshold the receive benefits. Districts with minority language populations close to the threshold were the most affected by differential privacy, with some classified correctly less than half of the time. This is because changing only a few individuals' data would change the ground truth of the threshold comparison, which is precisely what differential privacy aims to protect. With legislative apportionment, fairness requires apportioning Congressional seats proportionally to each district's population. This cannot be achieved exactly, even without noise, because the number of representatives in a district must be integral. Adding small amounts of noise for privacy actually improves fairness in expectation, because randomization can reduce (on average) the deviation from the quota caused by the integrality constraint. However, ex-post, any realized outcome cannot achieve the expected (fractional) allocation, so this may be an unsatisfying guarantee in practice.

[CDMS21] empirically measured the accuracy impact of differential privacy in Census data for the task of *redistricting*, or redrawing the boundaries of voting districts based on population counts. They focused on the actual differentially private algorithm used by the U.S. Census Bureau, named TopDown, as well as a simplified and easier-to-analyze variant named ToyDown, which both involve geographical hierarchical constraints relating population counts of larger regions (e.g., states) to the smaller subdivisions contained therein (e.g., county, track, block).

10. Importantly, the Census Public Use Microdata is *not* the ground truth of the American population distribution, as these data already include disclosure avoidance measures, including (prior to 2020) *swapping*, where certain database entries have been swapped. Swapping has been shown to provide insufficient privacy protections, as substantial fractions of the true database can be reconstructed exactly, even after swapping has been applied [Abo21]. However, for the sake of analysis, the authors required a dataset that could be treated as ground truth, against which to measure the performance of differentially private algorithms.

These algorithms are first used to privatize the population counts in every relevant geographical region of the country, and then these private population counts are used to construct new voting districts which have approximately equal populations.

Among other results on TopDown and ToyDown—including the impact of privacy budget allocation across geographical hierarchy and the impact of the requiring geometric structure when constructing voting districts—the authors consider the robustness of linear regression to predict voting outcomes of each precinct based on demographics. This regression approach is commonly used to asses racial polarization of precincts for legal enforcement of the Voting Rights Act. The authors showed that standard linear regression based on privatized data introduces significant bias in the estimation parameters. However, simple post-processing of the private data, such as removing extremely small precincts or weighting precincts by population removes virtually all of this error.

Overall these results suggest that while differential privacy has been successful in providing accuracy *in aggregate*—e.g., additive accuracy between the ground truth value of a population statistic and the result of a DP algorithm to estimate that statistic—there has been less attention to accuracy in terms of disparity. This work highlights that more attention is needed to accuracy *as a fairness measure* of private algorithms—e.g., accurate allocation of funding or accurate classification of the existence of a minority population. A natural approach is the joint design of differentially private algorithms and decision-making processes to explicitly accommodate the noise from DP algorithms. For example, shifting down the threshold used in the comparison test for voting rights benefits would ensure that populations close to the true threshold are still likely to receive the benefits they deserve. While this might slightly increase costs of printing voting materials (i.e., increase false positive rate), it would substantially reduce disenfranchisement of minority voters (i.e., decrease false negative rate). More research is needed to develop richer tools for the growing number of practical use-cases where downstream decisions are made based on the results of differentially private data analyses.

17.5.3 Fairness Composition in Complex Systems

Differential privacy is known to enjoy *composition*, meaning that when multiple private analyses are performed, the privacy of the system is guaranteed by the privacy of its components. Unfortunately, [DI19] showed that fairness does not share this same property, and classifiers which are fair in isolation are not guaranteed to compose into fair systems. Additionally, unfair components can be used to build a system that is fair overall.

[DI19] give several practical examples of where fairness composition may fail:

- Settings where multiple tasks compete for individuals, e.g., online advertising for employment where bids must be individually fair with respect to job qualification. Due to competition with other advertisers, the cost to display the same ad may be different for similarly qualified individuals, so any fixed bid for a given qualification level will yield different outcomes between individuals who should receive similar treatment.
- Composing fair systems with other systems that legitimately differentiate based on the protected attribute, e.g., a diaper advertiser bids higher on women who are mothers, and an employment advertiser bids fairly across genders. In equilibrium, women who are mothers will see the diaper ad, while women without children will see the employment ad. This is distinct from the previous example because the (non-protected) diaper ad advertiser is free to bid "unfairly" with respect to parental status.
- Classifiers built as functions of fairly obtained values, e.g., getting into College A *OR* College B *AND* receiving financial aid, when each admissions decision is fair. In general, functional compositions of fair components are not guaranteed to be fair, although in certain cases, composition using only *OR*s can be fair.
- Approximate fairness with feedback loops, e.g., admission into a good high school will improve the chance of college admission, when each admissions decision is approximately fair. Slight unfairness early in the process may compound to have a large effect on fairness of final outcomes.
- Settings where each individual's classification is dependent on the classifications of others, e.g., sequentially interviewing candidates. Selecting an earlier candidate precludes later candidates from even being considered, even if each candidate is fairly evaluated.

While [DI19] give algorithmic solutions for achieving fairness in each example above, these results highlight that fairness of a system cannot be guaranteed simply by making fair decisions independently at each step of the process. Practitioners wishing to design fair complex systems must analyze and account for interactions between system components to ensure fair outcomes overall.

17.6 Concluding Remarks

This chapter has summarized the relationships between differential privacy and three different notions of algorithmic fairness: individual fairness, group fairness, and multi-group fairness. Each of these relationships takes on a very different

flavor. The core definition of individual fairness is inspired by differential privacy, and techniques for achieving individual fairness include differentially private algorithms. Group fairness and multi-group fairness, on the other hand, are not inherently linked to the definition of differential privacy, but one can naturally ask whether these objectives can be achieved simultaneously with differential privacy. For both group and multi-group fairness, we see that this is possible. We also observed in Section 17.5 the potential negative impacts of differential privacy on fair outcomes if no explicit fairness interventions are taken. This suggests there is value to be gained from jointly considering privacy and fairness in learning systems.

Before these tools can be brought to bear to achieve privacy and fairness in machine learning systems in practice, there are a number of surrounding questions that must be addressed first. The most obvious is developing an understanding of when each fairness definition is appropriate for use. Naturally, it will depend on the application domain and the real-world fairness considerations that are being modeled in the algorithmic problem. As with any application of differential privacy, there is also the question of an appropriate choice of ε. The results surveyed in this chapter characterize the trade-off between privacy and fairness—as well as other desiderata such as accuracy or computational efficiency—as a function of ε, the sample size n, and other relevant problem-specific parameters. These insights should prove valuable to practitioners who wish to build and use systems that satisfy these desiderata, as well as to policymakers and regulators when developing new legislation to govern the use of AI systems.

Acknowledgements

The author would like to thank Michael Kim for extensive conversations and references on multi-group fairness, as well as comments on an earlier draft. Much of Section 4 and its framing is due to those conversations. The author would also like to thank Juba Ziani and Moon Duchin for conversations and references that were helpful in preparing this chapter.

References

[Abo21] J. Abowd. Third Declaration of John M. Abowd, Appendix B, Case 1:21-cv-01361-ABJ. https://www2.census.gov/about/policies/foia/records/disclosure-avoidance/17-1-abowd-decl-3.pdf [Last accessed 07/26/22]. Nov. 2021 (cit. on p. 584).

[Ala19] D. Alabi. "The Cost of a Reductions Approach to Private Fair Optimization". arXiv pre-print 1906.613. 2019 (cit. on p. 572).

[BPS19] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. "Differential Privacy Has Disparate Impact on Model Accuracy". In: Proceedings of the 33rd Conference on Neural Information Processing Systems. NeurIPS '19. 2019 (cit. on pp. 582, 583).

[CDMS21] A. Cohen, M. Duchin, J. Matthews, and B. Suwal. "Census Top-Down: The Impacts of Differential Privacy on Redistricting". In: Proceedings of the 2nd Symposium on Foundations of Responsible Computing. Ed. by K. Ligett and S. Gupta. FORC '21. 2021, 5:1–22 (cit. on pp. 583, 584).

[CGKM19] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern. "On the compatibility of privacy and fairness". In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization. FairUMAP '19. 2019, pp. 309–315 (cit. on pp. 567, 569–571).

[DI19] C. Dwork and C. Ilvento. "Fairness Under Composition". In: Proceedings of the 10th Innovations in Theoretical Computer Science. ITCS '19. 2019, 33:1–20 (cit. on pp. 574, 585, 586).

[Dwo+12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. "Fairness through awareness". In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ITCS '12. 2012, pp. 214–226 (cit. on pp. 559–561, 564, 565).

[FS96] Y. Freund and R. E. Schapire. "Game theory, on-line prediction and boosting". In: Proceedings of the Ninth Annual Conference on Computational Learning Theory. COLT '96. 1996, pp. 325–332 (cit. on pp. 569, 571, 572, 577, 579).

[Gop+22] P. Gopalan, A. T. Kalai, O. Reingold, V. Sharan, and U. Wieder. "Omnipredictors". In: Proceedings of the 13th Innovations in Theoretical Computer Science Conference. ITCS '22. 2022 (cit. on p. 578).

[HKRR18] U. Hebert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum. "Calibration for the (Computationally-Identifiable) Masses". In: Proceedings of the 35th International Conference on Machine Learning. ICML '18. 2018, pp. 1939–1948 (cit. on pp. 576, 577, 579, 580).

[HPS16] M. Hardt, E. Price, and N. Srebro. "Equality of Opportunity in Supervised Learning". In: Advances in Neural Information Processing Systems. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf (cit. on p. 568).

[Ilv20] C. Ilvento. "Metric learning for individual fairness". In: Proceedings of the Symposium on Foundations of Responsible Computing. FORC '20. 2020 (cit. on pp. 560, 564, 565).

[Jag+19] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman. "Differentially private fair learning". In: Proceedings of the International Conference on Machine Learning. ICML '19. PMLR. 2019, pp. 3000–3008 (cit. on pp. 567, 569–571).

[Jun+20] C. Jung, M. Kearns, S. Neel, A. Roth, L. Stapleton, and Z. S. Wu. "An Algorithmic Framework for Fairness Elicitation". In: 2nd Symposium on Foundations of Responsible Computing. FORC '20 2. Leibniz International Proceedings in Informatics, 2020, pp. 1–19 (cit. on pp. 560, 564, 565).

[Kim20] M. P. Kim. "A Complexity-Theoretic Perspective on Fairness". PhD thesis. Stanford University, 2020 (cit. on pp. 580, 581).

[KMR17] J. Kleinberg, S. Mullainathan, and M. Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: 8th Innovations in Theoretical Computer Science Conference. ITCS '17. 2017, 43:1–43:23 (cit. on pp. 566, 573).

[KNRW18] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness". In: Proceedings of the 35th International Conference on Machine Learning. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2564–2572 (cit. on pp. 570, 571).

[KRS19] M. Kearns, A. Roth, and S. Sharifi-Malvajerdi. "Average Individual Fairness: Algorithms, Generalization and Experiments". In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019 (cit. on p. 580).

[Mit+21] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. "Algorithmic Fairness: Choices, Assumptions, and Definitions". In: Annual Review of Statistics and Its Application 8.1 (2021), pp. 141–163 (cit. on pp. 566, 572).

[Nar18] A. Narayanan. 21 fairness definitions and their politics. Tutorial at Conference on Fairness, Accountability, and Transparency. 2018. URL: https://www.youtube.com/watch?v=jIXIuYdnyyk (cit. on pp. 566, 572).

[Puj+20] D. Pujol, R. McKenna, S. Kuppam, M. Hay, A. Machanavajjhala, and G. Miklau. "Fair Decision Making Using Privacy-Protected Data". In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20. 2020, pp. 189–199 (cit. on p. 583).

[RY18] G. N. Rothblum and G. Yona. "Probably Approximately Metric-Fair Learning". In: Proceedings of the International Conference on Machine Learning. ICML '18. 2018 (cit. on pp. 560, 565).

[SSV03] A. Sandroni, R. Smorodinsky, and R. V. Vohra. "Calibration with many checking rules". In: Mathematics of Operations Research 28.1 (2003), pp. 141–153 (cit. on p. 575).

[TDF21] C. Tran, M. Dinh, and F. Fioretto. "Differentially Private Empirical Risk Minimization under the Fairness Lens". In: Advances in Neural Information Processing Systems. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 27555–27565 (cit. on p. 583).