Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey

Ferdinando Fioretto[1] , **Cuong Tran**[1] , **Pascal Van Hentenryck**[2] and **Keyu Zhu**[2]
[1]Syracuse University
[2]Georgia Institute of Technology
{ffiorett, cutran}@syr.edu, pvh@isye.gatech.edu, kzhu67@gatech.edu

Abstract

This paper surveys recent work in the intersection of differential privacy (DP) and fairness. It reviews the conditions under which privacy and fairness may have aligned or contrasting goals, analyzes how and why DP may exacerbate bias and unfairness in decision problems and learning tasks, and describes available mitigation measures for the fairness issues arising in DP systems. The survey provides a unified understanding of the main challenges and potential risks arising when deploying privacy-preserving machine-learning or decisions-making tasks under a fairness lens.

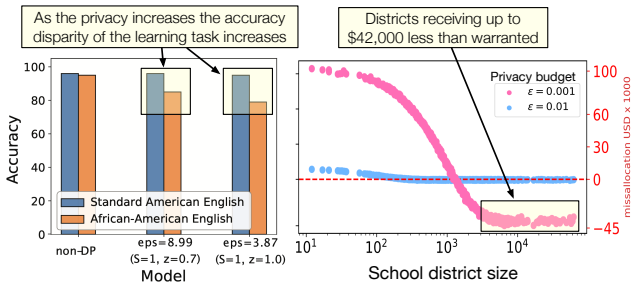

Figure 1: Left: Disparities arising in DP sentiment analysis tasks (image from [Bagdasaryan *et al.*, 2019]). Right: Disparity arising in fund allocations to school districts (image from [Tran *et al.*, 2021d]).

1 Introduction

The availability of large datasets and computational resources has driven significant progress in Artificial Intelligence (AI) and, especially, Machine Learning (ML). These advances have rendered AI systems instrumental for many decision making and policy operations involving individuals: they include assistance in legal decisions, lending, and hiring, as well determinations of resources and benefits, all of which have profound social and economic impacts. While data-driven systems have been successful in an increasing number of tasks, the use of rich datasets, combined with the adoption of black-box algorithms, has sparked concerns about how these systems operate. How much information these systems leak about the individuals whose data is used as input and how they handle biases and fairness issues are two of these critical concerns.

Differential Privacy (DP) [Dwork *et al.*, 2006] has become the paradigm of choice for protecting data privacy and its deployments are also growing at a fast rate. These include several data products related with the 2020 release of the US. Census Bureau [Abowd, 2018], and by Google [Aktay *et al.*, 2020], Facebook [Herdagdelen *et al.*, 2020], and Apple [Team, 2017]. DP is appealing as it bounds the risks of disclosing sensitive information for individuals participating in a computation. However, the process adopted by a DP algorithm to ensure the privacy guarantees involves calibrated perturbations, which inevitably introduce errors to the outputs of the task at hand. More importantly, it has been shown that these errors may affect different groups of individuals differently. An example of these effects are reported in Figure 1 (left), which illustrates that a DP learning model affects the accuracy of the minority group (African-American) more than it does the majority group in a sentiment analysis of Tweets [Bagdasaryan *et al.*, 2019]. Similar observations were reported in decision tasks (Figure 1, right) in which privacy-preserving census data is used to allocate funds to school districts [Pujol *et al.*, 2020; Tran *et al.*, 2021d]. The illustration shows that, under a privacy-preserving allocation scheme, some school districts may systematically receive considerably less money than what would be warranted otherwise.

These effects can have significant societal and economic impacts on the involved individuals: classification errors may penalize some groups over others in important determinations, including criminal assessment, landing, and hiring, or can result in disparities regarding the allocation of critical funds, benefits, and therapeutics. These fairness issues in DP settings are receiving increasing attention, but a complete understanding of why they arise is still limited. For example, it is often believed that post-processing the output of a differential private data-release mechanism may introduce bias and reduce errors but the underlying phenomena have only recently started to be studied in detail. Furthermore, in privacy-preserving learning tasks, it is (often incorrectly) believed that disparate impacts are caused by the presence of unbalanced data. *It is the goal of this survey to demystify some common beliefs about the interaction between differential privacy and fairness, and provide a critical review of the state of knowledge in this important area.* The survey fo-

cuses on two key privacy-preserving processes: *downstream decisions tasks*, in which a privacy-preserving version of a sensitive dataset is used to allocate resources or grant benefits, and *learning tasks*, in which a learning model is rendered differentially private.

2 Preliminaries

This section reviews the notion of differential privacy and compares some key fairness concepts adopted in this survey.

Differential Privacy (DP) [Dwork *et al.*, 2006] is a rigorous privacy notion that characterizes the amount of information of an individual's data being disclosed in a computation. A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ with domain $\mathcal{X}$ and range $\mathcal{Y}$ satisfies (ϵ, δ)-*differential privacy* if, for any output $y \in \mathcal{Y}$ and datasets $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ differing by at most one entry,

$$\Pr[\mathcal{M}(\boldsymbol{x}) = y] \leq \exp(\epsilon) \Pr[\mathcal{M}(\boldsymbol{x}') = y] + \delta. \quad (1)$$

Intuitively, DP states that outputs to the privacy-preserving mechanism are returned with a similar probability regardless of whether the dataset includes a specific individual. Parameter $\epsilon > 0$ describes the *privacy loss* of the mechanism, with values close to 0 denoting strong privacy. When $\delta = 0$, mechanism $\mathcal{M}$ is said to be ϵ-differentially private. Differential privacy satisfies several important properties. Notably, *composability* ensures that a combination of DP mechanisms preserves differential privacy and *post-processing immunity* ensures that privacy guarantees are preserved by arbitrary data-independent post-processing steps [Dwork and Roth, 2013].

Fairness. This survey focuses on two main fairness notions: *individual* and *group* fairness. Individual fairness [Dwork *et al.*, 2012] claims that *similar individuals should be treated similarly*. For a mechanism $\mathcal{M}$ mapping inputs $\boldsymbol{x} \in \mathcal{X}$ to outputs $y \in \mathcal{Y}$, individual fairness is satisfied when for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$:

$$d_{\mathcal{Y}}(\mathcal{M}(\boldsymbol{x}), \mathcal{M}(\boldsymbol{x}')) \leq d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}'), \quad (2)$$

where $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and $d_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, are distance metrics over pairs of inputs and outputs, respectively. When condition (2) holds, mechanism $\mathcal{M}$ is said to satisfy a $(d_{\mathcal{X}}, d_{\mathcal{Y}})$-Lipschitz condition. An obvious drawback of individual fairness is its requirement of problem-specific distance metrics, which may not be easy to design.

Group fairness, in contrast, requires some statistical property of any protected group of individuals (e.g., a defined by gender or race) to be similar to that of the whole population. Examples of commonly adopted group fairness notions are *demographic parity*, which is satisfied when the outputs of a predictor $\mathcal{M}$ are statistically independent of the protected group attribute [Dwork *et al.*, 2012], *equal opportunity*, which is satisfied when $\mathcal{M}$'s predictions are conditionally independent of the protected group attribute for a given label [Hardt *et al.*, 2016], and *accuracy parity*, which is satisfied when $\mathcal{M}$'s miss-classification rate is conditionally independent of the protected group attribute [Zhao and Gordon, 2019]. A connection between group fairness and individual fairness was presented by Dwork et al. [2012], who showed that if a model is individually fair and if the Earthmover distance across the protected groups data is sufficiently small, then such model also satisfies demographic parity.

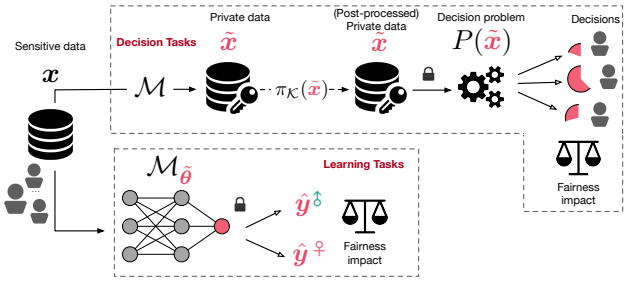

Figure 2: Setting analyzed in this survey.

3 Settings

The focus of the survey is to shed light on the disproportionate effects induced by a DP mechanism $\mathcal{M}$ on the outputs of some task of interest. The considered mechanisms process inputs $\boldsymbol{x} \in \mathcal{X}$ of n entries, containing sensitive information, such as the individuals' ethnicity, salary, gender, and geographic locations. Within this setting, the survey focuses on *decision tasks* and *learning tasks*, whose schematic illustrations are shown in Figure 2.

Decision Tasks. This setting considers data-release mechanisms $\mathcal{M}$ producing a privacy-preserving counterpart $\tilde{\boldsymbol{x}}$ of $\boldsymbol{x}$. Then the DP dataset $\tilde{\boldsymbol{x}}$ is used as the input to a decision problem $P : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$. For instance, P may describe an allotment of funds to school districts. This setting is widely adopted in several data-release tasks, including census applications and allocation of renewable energy resources in energy markets. Mechanism $\mathcal{M}$ may also apply a post-processing step $\pi_{\mathcal{K}}$ to restrict the randomized output $\tilde{\boldsymbol{x}}$ to be within a feasible region $\mathcal{K}$, e.g., to guarantee non-negativity of the released data. The focus of this task is to study the effects of a DP data-release mechanism $\mathcal{M}$ to the outcomes of problem P in relation to the fairness of the decisions. Because random noise is added to the original dataset $\boldsymbol{x}$, the output $P(\tilde{\boldsymbol{x}})$ incurs some error. A quantification of the disparate impact of this error among the problem entries is often measured through the bias of problem P for some entry $i \in [n]$,

$$B_P^i(\mathcal{M}, \boldsymbol{x}) = \mathbb{E}_{\tilde{\boldsymbol{x}} \sim \mathcal{M}(\boldsymbol{x})} [P_i(\tilde{\boldsymbol{x}})] - P_i(\boldsymbol{x}), \quad (3)$$

where P_i is used to denote the program computing the output associated with entity $i \in [n]$. This bias characterizes the distance between the expected privacy-preserving decision and the decision obtained on the real data (ground truth). In this context, the fairness analysis attempts to bound the maximal difference in bias among any pairs of entries: $\max_{i,j \in [n]} |B_P^i(\mathcal{M}, \boldsymbol{x}) - B_P^j(\mathcal{M}, \boldsymbol{x})|$.

Learning Tasks. In this second setting, $\mathcal{M}_{\tilde{\theta}}$ is a classifier parametrizded by vector $\tilde{\theta}$ that protects the disclosure of the individuals in $\boldsymbol{x}$ and the focus is to analyze the fairness impact of privacy on different groups of individuals. The elements of $\boldsymbol{x}$ are data points (x, a, y) where $x \in \mathcal{X}$[1] is a feature vector, $a \in \mathcal{A}$ is a protected group attribute, and $y \in \mathcal{Y}$ is a label. The model quality is measured by a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, and the problem is to minimize the empirical risk function:

[1] Used here to denote the feature set, slightly abusing notation.

$\min_\theta \mathcal{L}(\theta; \boldsymbol{x}) = \mathbb{E}_{(x,a,y)}[\ell(\mathcal{M}_\theta(x), y)]$. Methods reviewed in this survey analyze the disparate impact of privacy on different groups of individuals either by measuring the deviation from a model to satisfy a notion of group fairness exactly or using the notion of *excessive risk* [Zhang *et al.*, 2017; Wang *et al.*, 2019]. The latter defines the difference between the private and non private risk functions:

$$R(\theta, \boldsymbol{x}_a) = \mathbb{E}_{\tilde{\theta}}\left[\mathcal{L}(\tilde{\theta}; \boldsymbol{x}_a)\right] - \mathcal{L}(\theta^*; \boldsymbol{x}_a), \tag{4}$$

where the expectation is defined over the randomness of the private mechanism, $\boldsymbol{x}_a$ denotes the subset of $\boldsymbol{x}$ containing exclusively samples whose group attribute is a, $\tilde{\theta}$ denotes the private model parameters, and $\theta^* = \operatorname{argmin}_\theta \mathcal{L}(\theta; \boldsymbol{x})$. In this context, (pure) fairness is achieved when there is no difference in excessive risk across all protected groups.

4 Privacy and Fairness: Friends or Foes?

While DP aims at rendering the participation of individuals indistinguishable to an observer who accesses the outputs of a computation, fairness attempts at equalizing properties of these outputs across different individuals. Thus, simultaneously achieving these two goals has received two contrasting views. The first sees privacy and fairness as *aligned* objectives while the second sees them as *contrasting* ones.

Contributions in the "aligned space" focus on studying conditions for which privacy and fairness can be achieved simultaneously. Notably, Dwork et al. [2012] shows that individual fairness is a generalization of differential privacy. To see why privacy and fairness may be achieved simultaneously, notice that a mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ satisfies ϵ-differential privacy when it is $(d_\mathcal{X}, d_\mathcal{Y})$-Lipschitz with

$$d_\mathcal{X}(\boldsymbol{x}, \boldsymbol{x}') = \epsilon|\boldsymbol{x}\Delta\boldsymbol{x}'|$$
$$d_\mathcal{Y}(\mathcal{M}(\boldsymbol{x}), \mathcal{M}(\boldsymbol{x}')) = \sup_{y\in\mathcal{Y}} \log\left(\frac{\Pr(\mathcal{M}(\boldsymbol{x}) = y)}{\Pr(\mathcal{M}(\boldsymbol{x}') = y)}\right),$$

where $\boldsymbol{x}\Delta\boldsymbol{x}'$ represents the set difference between two inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$ of $\mathcal{X}$. Thus, DP mechanisms also ensure individual fairness, as long as $d_\mathcal{X}$ and $d_\mathcal{Y}$ are defined as above. Similarly, Mahdi et al. [2021] shows that, in candidate selection problems, the use of a DP exponential mechanism [McSherry and Talwar, 2007] produces fair selections when the data satisfies some restrictions concerning key properties (average and variance of the qualification scores) of each group.

The second line of works views privacy and fairness as contrasting goals. Notably, it has been observed that the outputs of DP classifiers may create or exacerbate disparate impacts among groups of individuals [Bagdasaryan *et al.*, 2019]. A similar phenomenon was also reported in important decision tasks that use DP census statistics as inputs [Pujol *et al.*, 2020]. These works typically adopt the notion of group fairness and impose no restrictions on the properties of the privacy-preserving mechanisms studied. The rest of the survey focuses on analyzing why these important observations arise and how can they be mitigated.

5 Why Privacy Impacts Fairness?

This section reviews the current understanding about why disparate impacts arise in two common privacy-preserving processes: decision tasks and learning tasks.

5.1 Decision Tasks

Consider first a data-release mechanism $\mathcal{M}$, which typically consists of two steps: First, noise drawn from a calibrated distribution is injected into the original data $\boldsymbol{x}$ to obtain a DP counterpart $\tilde{\boldsymbol{x}}$. This process, however, may fundamentally affect some important properties of the original data. For example, if $\boldsymbol{x}$ is a vector of counts enumerating individuals living in different regions, its privacy-preserving version $\tilde{\boldsymbol{x}}$ may not satisfy non-negativity conditions. Thus, a post-processing step $\pi_\mathcal{K}$ is applied to $\tilde{\boldsymbol{x}}$ to redistribute the noisy values in a way that the resulting outputs $\pi_\mathcal{K}(\tilde{\boldsymbol{x}})$ satisfy the desired data-independent constraints $\mathcal{K}$. Second, the released data $\tilde{\boldsymbol{x}}$ is used as input to a decision problem P. This pipeline is shown in Figure 2 (top). The goal of this section is to characterize the disparity in errors induced by mechanism $\mathcal{M}$ on the final decisions $P(\tilde{\boldsymbol{x}})$.

The negative impacts of privacy towards fairness in decision tasks were first observed by Pujol et al. [2020]. The authors noticed that the use of privacy-preserving census data to allocate funds to school district produces unbalanced allocation errors, with some school districts systematically receiving more (or less) than what warranted, as illustrated in Figure 1 (right). A similar behavior was also observed in other census-motivated decision tasks, including determining whether a jurisdiction qualifies for providing minority language assistance during an election, and apportionment of legislative representatives.

These empirical observations were later attributed to two main factors: (1) the "shape" of the decision problem P [Tran *et al.*, 2021d] and (2) the presence of non-negativity constraints in post-processing steps [Zhu *et al.*, 2021]. The survey reviews next these two factors in details.

Shape of the Decision Problem. Note that private data is often calibrated with unbiased noise, such as Laplacian noise in the Laplace mechanism, for privacy protection. In such contexts Tran et al. [2021d] showed that a decision problem that applies a linear transform of its input yields an unbiased outcome with respect to the true outcome. However, non-linearities in the decision problem are likely to generate non-zero biases with discrepancies among entities, which results in fairness issues. In more details, when P_i is at least twice differentiable, the problem bias can be approximated as

$$\begin{aligned} B^i_P(\mathcal{M}, \boldsymbol{x}) &= \mathbb{E}[P_i(\tilde{\boldsymbol{x}} = \boldsymbol{x} + \eta)] - P_i(\boldsymbol{x}) \\ &\approx \frac{1}{2}\boldsymbol{H}P_i(\boldsymbol{x}) \times \mathrm{Var}[\eta] \end{aligned} \tag{5}$$

where $\boldsymbol{H}P_i(\cdot)$ denotes the Hessian of problem P_i. The approximation above uses a Taylor expansion of the private problem $P_i(\boldsymbol{x}+\eta)$ and the linearity of expectations, with η a random variable following some symmetric distribution. The bias B^i_P can thus be approximated by an expression involving the local curvature of the problem P_i and the variance of the noisy input (which depends on the privacy loss ϵ). In

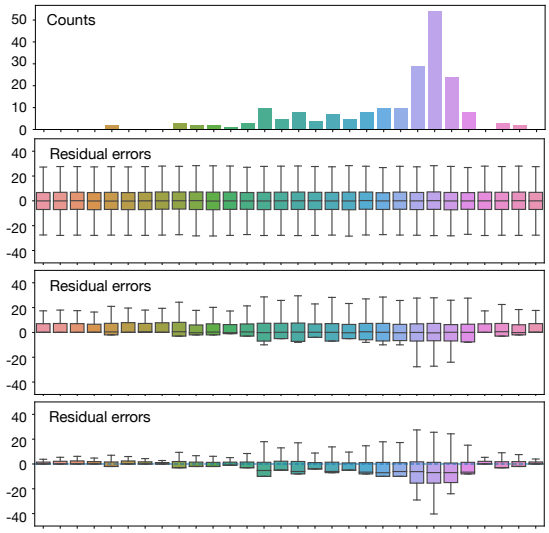

Figure 3: Bias and variance in DP post-processing.

turn, fairness violations (the maximal difference of the bias between any two entries) can be bounded whenever the problem local curvature is constant across entities, since the variance is also constant and bounded.

Observe that the fairness violations are controlled by both the privacy loss value ϵ (appearing in the variance term) and the shape of the decision problem (appearing in the Hessian term). Tight privacy requirements (small ϵ values) or non-linearities in the decision problem may lead to large disparate impacts. An important conclusion of this result is that using DP to generate private inputs of decision problems commonly adopted to make policy determinations *will necessarily introduce fairness issues, despite the noise being unbiased*.

Notice that the analysis above holds for problems with continuous support. In the case of *Boolean decision functions*, e.g., $P_i : \mathcal{X} \to \{0, 1\}$, it was shown that disparate impacts may be exacerbated when multiple decision functions are composed [Tran *et al.*, 2021d]. This is of particular interest in policy decisions like those used to determine whether a jurisdiction qualifies for a particular benefit, such as the minority language voting right problem [Pujol *et al.*, 2020] where thresholding Boolean functions are "concatenated" using logical connectors. In a nutshell, the result shows that composing two decision problems with fairness violations bounded by values α_1 and α_2, respectively, produces a fairness violation bound $\alpha > \max(\alpha_1, \alpha_2)$.

Impact of Post-processing. Post-processing immunity is a fundamental property of differential privacy which is routinely applied in many applications, including census data [Abowd, 2018], energy systems [Fioretto *et al.*, 2020b], and mobility [He *et al.*, 2015; Fioretto *et al.*, 2018]. Notably, when the feasible region is convex, a largely adopted class of post-processing functions, called *projections*, is guaranteed to improve accuracy [Hay *et al.*, 2010; Fioretto *et al.*, 2021].

While post-processing is often used to reduce errors, this step can also introduce bias and fairness issues, as illustrated by Zhu et al. [2021] and McGlinchey et al. [2020]. The issue is depicted in Figure 3. It displays a histogram of population counts (top row) and the distribution of the residual errors $\tilde{\boldsymbol{x}} - \boldsymbol{x}$, where $\tilde{\boldsymbol{x}}$ is obtained by the application of Laplace noise on the true counts (second row). The third row displays the residual errors when applying post-processing step

$$\pi_{\geq 0} \coloneqq \max(\mathbf{0}, \tilde{\boldsymbol{x}})$$

to enforce non-negativity. Finally, the fourth row is obtained by a constrained projection method

$$\pi_{\mathcal{K}_S} \coloneqq \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{K}_S} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2, \ \mathcal{K}_S = \{\boldsymbol{v} \in \mathbb{R}^n \mid \sum_i v_i = \tilde{S}, v_i \geq 0\},$$

that enforces a linear constraint imposing that the sum of the projected counts $\boldsymbol{v}$ be equal to a noisy sum $\tilde{S} = (\sum_i x_i) + \eta$ with η being some appropriately selected noise. All of the above private methods achieve the same privacy loss. Note that the application of Laplace noise does not introduce bias: all the outputs have the same residual errors. However, even the simple non-negative post-processing step produces bias and different error variances across the problem entries [McGlinchey and Mason, 2020]. The more complex mechanism $\pi_{\mathcal{K}_S}$ further exacerbates the biases and variance differences as showed by Zhu et al. [2022]. Additionally, Zhu et al. [2021] showed that for the more general constraint spaces $\mathcal{K} = \{\boldsymbol{x} \mid A\boldsymbol{x} \leq b, \boldsymbol{x} \geq 0\}$, the solution $\pi_{\mathcal{K}}(\tilde{\boldsymbol{x}})$ is an unbiased estimator of $\boldsymbol{x}$ when the non-negative constraint $\boldsymbol{x} \geq \mathbf{0}$ is ignored. However, when incorporating this non-negativity constraint, the optimal solution $\pi_{\mathcal{K}}(\tilde{\boldsymbol{x}})$ deviates from $\boldsymbol{x}$ statistically and, thus, resulting in non-zero bias.

5.2 Learning Tasks

Consider now the disparate impacts arising in private learning tasks. These effects were first studied in the context of equal opportunity by Cummings et al. [2019]. They show that it is impossible to achieve pure fairness, i.e., $\Pr[\mathcal{M}(\boldsymbol{x}) = \hat{y} \mid a, y] = \Pr[\mathcal{M}(\boldsymbol{x}) = \hat{y} \mid y]$, for all $a \in \mathcal{A}$, when using an ϵ-DP classifier. This result relies on the observation that a classifier $\mathcal{M}$ cannot be perfectly fair for both a dataset $\boldsymbol{x}$ and a neighbor dataset $\boldsymbol{x}'$ differing from $\boldsymbol{x}$ by adding or removing one sample. However, a relaxed fairness goal, i.e., $0 < |\Pr[\mathcal{M}(\boldsymbol{x}) = \hat{y} \mid a, y] - \Pr[\mathcal{M}(\boldsymbol{x}) = \hat{y} \mid y]|$ can be achieved by using the exponential mechanism, which satisfies ϵ-DP [Cummings *et al.*, 2019].

The effects of privately training deep learning models to the accuracy parity was first observed by Bagdasaryan et al. [2019]. The authors studied the disparities induced by DP Stochastic Gradient Descent (DP-SGD) [Abadi *et al.*, 2016], the de-facto standard algorithm used to train deep learning models privately. They observed that the accuracy of the minority group was disproportionately impacted by the private training. These observations were validated on several vision and natural language processing tasks and in both a centralized and federated setting. The authors postulated that the size of a protected group would play a crucial role to the exacerbation of the disparate impacts in private training. This behavior was also observed in a further empirical study which compared DP-SGD and PATE [Papernot *et al.*, 2018], a popular semi-supervised DP learning framework. Therein, the authors reported PATE to cause milder disparate impacts

when compared to DP-SGD under similar privacy constraints [Uniyal *et al.*, 2021]. The hypothesis considering the size of a protected group as a predominant factor inducing the disparate impacts in private training was challenged by Farrand et al. [2020]. The authors showed that the privacy preserving models can introduce substantial fairness issues even when slightly imbalanced datasets are considered.

More recently, Tran et al. [2021a; 2021b] show that group sizes may indeed not be a predominant factor to explain the disparate impacts observed in private training. The authors report two main factors contributing to these effects: **(1)** the properties of the training data and **(2)** the model's characteristics. The survey reviews these factors next.

Properties of the Training Data. As observed by Tran et al. [2021a], *input norms* and *distance to decision boundary* are two key characteristics of the data connected with exacerbating the disparate impacts of private learning tasks. First, by using an analysis analogous to that used in Equation (5), it was shown that, when training convex models using output perturbation–which adds Gaussian noise to the output of the optimal models parameters–groups of samples associated with large Hessian losses $\boldsymbol{H}\ell(\theta; \boldsymbol{x}_a)$ can be penalized more than those associated with small Hessian losses. In turn, the authors show that groups with large input norms (often observed at the tail of the data distribution) may lead to large Hessian loss values. The observation that samples at tail of the data distribution are often penalized more than others was also reported by Bagdasaryan et al. [2019] and Suriyakumar et al. [2021]. Additionally, Tran et al. [2021a] show that the distance of a sample to the model decision boundary is also connected to the Hessian values. Samples which are near (far) to the decision boundary are less (more) tolerant to perturbations induced by the DP algorithm. This is intuitive, since perturbing the model parameters is more likely to impact samples which are close to the decision boundary. Similar observations were also reported in the context of DP-SGD [Tran *et al.*, 2021a] and PATE [Tran *et al.*, 2021b].

Model Characteristics. In addition to the data properties, the characteristics of DP learning mechanisms have also been found connected with the disparate impacts of the private models. For example, at each training iteration, DP-SGD operates by computing the gradients for each data sample in a random mini-batch, clipping their L2-norm, adding noise to ensure privacy, and computing the average. The two key characteristics of DP-SGD are clipping the gradients whose L2 norm exceeds a given bound C and perturbing the averaged clipped gradients with Gaussian noise. As shown in [Tran *et al.*, 2021a], both factors exacerbate unfairness in the private predictions of DP-SGD. When different groups of individuals produce updates with large differences in magnitude of gradients norms, and when such values exceed the clipping bound C, then gradient clipping induces dissimilar information losses in these groups, thus penalizing those groups with larger gradients. This aspect was also observed in [Xu *et al.*, 2021]. Additionally, the process of adding noise in DP-SGD is shown to produce an effect similar to that produced by the output perturbation reviewed above: The groups with larger Hessian losses defined on their samples tend to have the larger disparate impacts [Tran *et al.*, 2021a].

Another important private ML framework is the *Private Aggregation of Teacher Ensembles (PATE)* [Papernot *et al.*, 2018]. It combines multiple learning models used as teachers for a student model that learns to predict an output chosen by noisy voting among the teachers. The resulting model satisfies differential privacy and has been shown effective in learning high quality private models in semisupervised settings [Malek Esmaeili *et al.*, 2021]. A key aspect of PATE is the scheme adopted by its teachers ensemble to privately predict labels which are used to train the student model. Tran et al. [2021b] showed that both the size of this ensemble and the confidence of the voting teachers are key factors in the analysis of the disparate impacts observed in this framework. Their analysis indicates that larger ensembles correspond to more robust predictions since the voting scheme becomes more consistent, given the noise added to guarantee privacy.

6 Mitigating Fairness in Private Tasks

Having discussed the reason why disparate impacts arise in differentially private decision making and learning tasks, the survey reviews next the strategies proposed in the literature to mitigate the fairness issues arising in these two settings.

6.1 Decision Tasks

Several solutions have recently been developed to reduce the disparate impacts arising in DP decision tasks, with particular focus on a class of census-motivated problems used to distribute funds or grants benefits to the problem entities. In particular, in the context of funds allocation to school district, Pujol et al. [2020] proposed a mechanism which distributes additional budget to targeted entities, so that, all of the entities receive at least what warranted in a non-private allocation, with high probability. A limitation of this strategy is that the resulting allocation does not necessarily ensures feasibility (e.g., the sum of the individual allocations may not match a preassigned budget). In the attempt to mitigate fairness issues for an analogous classes of problems, Tran et al. [2021d] proposed to design a proxy problem that closely approximates the original decision task but admits bounded fairness. In particular, the authors observe that a linear approximation for an important class of allocation problems can be obtained when additional aggregated data can be released.

Solutions to mitigate the disparate effects of post-processing have also been proposed. In particular, Zhu et al. [2022] analyzed the fairness impact of projection mechanisms on a simplex and proposed a near-optimal projection operator which meets the feasibility requirements of allotment problems while providing substantial improvements in terms of disparate impact under different fairness metrics.

6.2 Learning Tasks

Mitigation strategies have also been designed in the context of learning tasks. Xu et al. [2019] and Ding et al. [2020] proposed versions of a fair and ϵ- and (ϵ, δ)-DP logistic regression classifiers [Zhang *et al.*, 2012]. Both works target demographic parity and use a functional mechanism–which approximates the objective function of the classifier by a polynomial and injects calibrated noise to its coefficients.

Most of the work attempting to mitigate the disparate impacts of DP in learning tasks has focused on the popular DP-SGD framework. DP-SGD does not restrict focus on convex loss functions rendering it an appealing framework for DP learning tasks. As observed in Section 5.2, individual gradient clipping is a key factor in exacerbating the disparate impacts. Thus, Xu et al. [2021] proposes to associate a different clipping bound to each protected group, so as to limit the effect of disproportionate gradient pruning for those groups of samples producing large gradients. This method has been shown to reduce the accuracy disparity across groups on tabular datasets. It was also noted that computing different clipping bounds leaks additional information when compared to classical DP-SGD, and thus requires larger perturbations. This causes an additional trade-off, since, as shown in Section 5.2, the noise magnitude is an important source of disparate impacts in DP-SGD. A different attempt uses early stopping [Zhang *et al.*, 2021], noting that the number of training iterations is a key factor to balance utility, privacy, and fairness. Importantly, this method relies on the availability of a public validation set. Finally, motivated by the observed differences in excessive risk across groups during private training (see Section 5.2), Tran et al. [2021a] suggest to add a fairness constraint to the empirical risk minimizer. The constraint equalizes the difference among the groups' excessive risks. This simple solution was shown to reducing the excessive risk differences among groups while retaining high accuracy.

The works above all protect each data sample in a dataset. The question regarding the necessity to protect exclusively the group attributes (e.g., gender or race) was first posed by Jagielski et al. [2019]. Under this more permissive privacy setting, the authors propose two algorithms that balance privacy and equalized odds impacts. The first is a DP version of the post-processing method of Hardt et al. [2016], which uses different decision thresholds for different groups to remove the disparate mistreatment.. The second is a DP version of the method suggested by Agarwal et al. [2018], which augments the loss function with a penalizer that accounts for the reported fairness violations. While innovative, these algorithms require very large privacy budgets, which is partly due to the use of advanced composition to derive the final privacy loss. In a similar setting, Mozannar et al. [2020] introduced a simple yet effective solution: It first applies randomized response to protect the sensitive group labels. Then, it uses the penalizer model introduced in [Agarwal *et al.*, 2018] to obtain a good tradeoff between privacy and fairness. Lastly, Tran et al. [2021c] study a private extension of a Dual Lagrangian framework applied to learning tasks [Fioretto *et al.*, 2020a], in which the primal and dual steps are computed privately and privacy computation relies on the moment accountant [Abadi *et al.*, 2016]. While good privacy/fairness trade-offs are reported, this method comes at a steeper computational cost due to the computations required by the private dual step.

In the context of semi-supervised teacher ensemble models, Tran et al. [2021b] observes that the voting process of the teacher ensemble is subject to robustness issues especially in low voting confidence regimes (small perturbations may significantly affect the result of the voting result). To mitigate this issue, the authors explore the use of soft-labels in the voting ensemble. Teachers using soft labels report confidence score associated with each target label, rather than reporting solely the label with the largest confidence. This additional information is carried at no additional privacy cost and it was shown helpful in achieving better privacy/fairness trade-offs.

Finally, some recent solution tackling privacy and fairness has also arisen in the context of federated learning. Notably, Abay et al. [2020] proposed several pre-processing and in-processing bias mitigation solutions to improve fairness without affecting data privacy. Finally, Padala et al. [2021] proposed a two-phase training step performed by each client. Clients first train a non-private model which maximizes accuracy while controlling the fairness violations. Then, they train a private model using DP-SGD to mimic the first, fair model. The updates obtained by this private model are thus broadcasted to an aggregator at each iteration.

7 Challenges and Research Directions

The current research at the intersection of differential privacy and fairness has shown promise in building solutions to realize more trustworthy systems. Furthermore, the analysis of the disparate impacts arising in several learning and decision tasks has paved the way to develop promising mitigating strategies. Despite these encouraging results, a number of challenges must be addressed to have a full understanding of the trade-offs between privacy, fairness, and accuracy. **(1)** The development of a unified theoretical framework to characterize and reason about fairness issues arising in general decision tasks is still missing. Of particular importance would be to capture the relation between the privacy loss values and the fairness violations resulting in both decision-making and learning settings. **(2)** While the current focus in the analysis of fairness in private ML tasks has focused on data and algorithmic properties, it has also been observed that batch-size and learning rate may affect the Hessian spectrum of a network classifier [Yao *et al.*, 2018]. These observations may suggest that fairness in private ML tasks may be impacted by key hyper-parameters, including batch size, learning rates, and the depth of neural networks. **(3)** Another aspect that has been observed repeatedly when connecting privacy and fairness is their link with model robustness. While this observation arises both in decision and in learning tasks, an understanding of this link is currently missing. **(4)** A further important direction is the study of the disparate impacts that may arise in algorithms and generative models producing private synthetic datasets as well the development of mitigation measures. **(5)** Finally, the development of software library to facilitate auditing fairness and bias issues in a private decision or learning task would be crucial to broaden the knowledge and adoption of these important issues.

Understanding the intricacies at the intersection of privacy, fairness, and accuracy will help shed light on the design of fairer ML systems and decision problems that use sensitive data. In turn, this will provide novel and unique perspectives for users and policymakers about the societal consequences of using differential privacy for critical processes, including predictions and decisions tasks.

Acknowledgments

This research is partially supported by NSF grant 2133169 and a Google Research Scholar Award. Its views and conclusions are those of the authors only.

References

[Abadi *et al.*, 2016] Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[Abay *et al.*, 2020] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv*, 2012.02447, 2020.

[Abowd, 2018] John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018.

[Agarwal *et al.*, 2018] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 2018.

[Aktay *et al.*, 2020] Ahmet Aktay, Shailesh Bavadekar, Gwen Cossoul, John Davis, Damien Desfontaines, Alex Fabrikant, Evgeniy Gabrilovich, Krishna Gadepalli, Bryant Gipson, Miguel Guevara, Chaitanya Kamath, Mansi Kansal, Ali Lange, Chinmoy Mandayam, Andrew Oplinger, Christopher Pluntke, Thomas Roessler, Arran Schlosberg, Tomer Shekel, Swapnil Vispute, Mia Vu, Gregory Wellenius, Brian Williams, and Royce J Wilson. Google covid-19 community mobility reports: Anonymization process description (version 1.1). *arXiv*, 2004.04145, 2020.

[Bagdasaryan *et al.*, 2019] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *International Conference on Neural Information Processing*, volume 32, 2019.

[Cummings *et al.*, 2019] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315, 2019.

[Ding *et al.*, 2020] Jiahao Ding, Xinyue Zhang, Xiaohuan Li, Junyi Wang, Rong Yu, and Miao Pan. Differentially private and fair classification via calibrated functional mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 622–629, 2020.

[Dwork and Roth, 2013] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

[Dwork *et al.*, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Farrand *et al.*, 2020] Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, pages 15–19, 2020.

[Fioretto *et al.*, 2018] Ferdinando Fioretto, Chansoo Lee, and Pascal Van Hentenryck. Constrained-based differential privacy for private mobility. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1405–1413, 2018.

[Fioretto *et al.*, 2020a] Ferdinando Fioretto, Pascal Van Hentenryck, Terrence WK Mak, Cuong Tran, Federico Baldo, and Michele Lombardi. Lagrangian duality for constrained deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2020.

[Fioretto *et al.*, 2020b] Ferdinando Fioretto, Terrence W.K. Mak, and Pascal Van Hentenryck. Differential privacy for power grid obfuscation. *IEEE Transactions on Smart Grid*, 11(2):1356–1366, March 2020.

[Fioretto *et al.*, 2021] Ferdinando Fioretto, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy of hierarchical census data: An optimization approach. *Artificial Intelligence*, 296:103475, 2021.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 2016.

[Hay *et al.*, 2010] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proc. VLDB Endow.*, 3(1):1021–1032, 2010.

[He *et al.*, 2015] Xi He, Graham Cormode, Ashwin Machanavajjhala, Cecilia M. Procopiuc, and Divesh Srivastava. Dpt: Differentially private trajectory synthesis using hierarchical reference systems. *Proc. VLDB Endow.*, 8(11):1154–1165, 2015.

[Herdagdelen *et al.*, 2020] Amaç Herdagdelen, Alex Dow, Bogdan State, Payman Mohassel, and Alex Pompe. Protecting privacy in facebook mobility data during the covid-19 response. t.ly/3-3W, 2020. Accessed: 2022-05-10.

[Jagielski *et al.*, 2019] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *International Conference on Machine Learning*. PMLR, 2019.

[Khalili *et al.*, 2021] Mohammad Mahdi Khalili, Xueru Zhang, Mahed Abroshan, and Somayeh Sojoudi. Improving fairness and privacy in selection problems. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

[Malek Esmaeili *et al.*, 2021] Mani Malek Esmaeili, Ilya Mironov, Karthik Prasad, Igor Shilov, and Florian Tramer. Antipodes of label differential privacy: Pate and alibi. *Advances in Neural Information Processing Systems*, 34, 2021.

[McGlinchey and Mason, 2020] Aisling McGlinchey and Oliver Mason. Observations on the bias of nonnegative mechanisms for differential privacy. *Foundations of Data Science*, 2(4):429, 2020.

[McSherry and Talwar, 2007] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[Mozannar *et al.*, 2020] Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning*. PMLR, 2020.

[Padala *et al.*, 2021] Manisha Padala, Sankarshan Damle, and Sujit Gujar. Federated learning meets fairness and differential privacy. In *International Conference on Neural Information Processing*, pages 692–699. Springer, 2021.

[Papernot *et al.*, 2018] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with PATE. In *6th International Conference on Learning Representations*, 2018.

[Pujol *et al.*, 2020] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *FAT* '20: Conference on Fairness, Accountability, and Transparency*, 2020.

[Suriyakumar *et al.*, 2021] Vinith M. Suriyakumar, Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. Chasing your long tails: Differentially private prediction in health care settings. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency,*, 2021.

[Team, 2017] Apple Differential Privacy Team. Learning with privacy at scale. *Apple Machine Learning Journal*, 1(8), 2017.

[Tran *et al.*, 2021a] Cuong Tran, My Dinh, and Ferdinando Fioretto. Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, 2021.

[Tran *et al.*, 2021b] Cuong Tran, My H Dinh, Kyle Beiter, and Ferdinando Fioretto. A fairness analysis on private aggregation of teacher ensembles. *arXiv*, 2109.08630, 2021.

[Tran *et al.*, 2021c] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

[Tran *et al.*, 2021d] Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. Decision making with differential privacy under a fairness lens. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[Uniyal *et al.*, 2021] Archit Uniyal, Rakshit Naidu, Sasikanth Kotti, Sahib Singh, Patrik Joslin Kenfack, Fatemehsadat Mireshghallah, and Andrew Trask. Dp-sgd vs pate: Which has less disparate impact on model accuracy? *arXiv*, 2106.12576, 2021.

[Wang *et al.*, 2019] Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[Xu *et al.*, 2019] Depeng Xu, Shuhan Yuan, and Xintao Wu. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 594–599, 2019.

[Xu *et al.*, 2021] Depeng Xu, Wei Du, and Xintao Wu. Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1924–1932, 2021.

[Yao *et al.*, 2018] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

[Zhang *et al.*, 2012] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: Regression analysis under differential privacy. *Proc. VLDB Endow.*, 2012.

[Zhang *et al.*, 2017] Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private ERM for smooth objectives. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

[Zhang *et al.*, 2021] Tao Zhang, Tianqing Zhu, Kun Gao, Wanlei Zhou, and S Yu Philip. Balancing learning model privacy, fairness, and accuracy with early stopping criteria. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[Zhao and Gordon, 2019] Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 2019.

[Zhu *et al.*, 2021] Keyu Zhu, Pascal Van Hentenryck, and Ferdinando Fioretto. Bias and variance of post-processing in differential privacy. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

[Zhu *et al.*, 2022] Keyu Zhu, Ferdinando Fioretto, and Pascal Van Hentenryck. Post-processing of differentially private data: A fairness perspective. *arXiv*, 2201.09425, 2022.