Privacy and Fairness in Machine Learning: A Survey

Sina Shaham, Arash Hajisafi*, Minh K. Quan*, Dinh C. Nguyen, Bhaskar Krishnamachari, Charith Peris, Gabriel Ghinita, Cyrus Shahabi, and Pubudu N. Pathirana

***Abstract*—Privacy and fairness are two crucial pillars of responsible Artificial Intelligence (AI) and trustworthy Machine Learning (ML). Each objective has been independently studied in the literature with the aim of reducing utility loss in achieving them. Despite the significant interest attracted from both academia and industry, there remains an immediate demand for more in-depth research to unravel how these two objectives can be simultaneously integrated into ML models. As opposed to well-accepted trade-offs, i.e., privacy-utility and fairness-utility, the interrelation between privacy and fairness is not well-understood. While some works suggest a trade-off between the two objective functions, there are others that demonstrate the alignment of these functions in certain scenarios. To fill this research gap, we provide a thorough review of privacy and fairness in ML, including supervised, unsupervised, semi-supervised, and reinforcement learning. After examining and consolidating the literature on both objectives, we present a holistic survey on the impact of privacy on fairness, the impact of fairness on privacy, existing architectures, their interaction in application domains, and algorithms that aim to achieve both objectives while minimizing the utility sacrificed. Finally, we identify research challenges in achieving privacy and fairness concurrently in ML, particularly focusing on large language models.**

***Impact Statement*—This extensive survey meticulously examines the intricate relationship between privacy and fairness in machine learning, revealing key alignments and trade-offs that have significant implications for the development of responsible AI systems. By consolidating terminology and identifying critical research gaps across various ML paradigms, this survey empowers researchers and practitioners to design and implement ML models that uphold both privacy and fairness, fostering trust and accountability in AI-driven decision-making processes. The insights presented in this survey are poised to catalyze future research aimed at developing innovative techniques that seamlessly integrate privacy and fairness considerations into the very fabric of ML algorithms.**

I. Introduction

The rapid expansion of big data, along with the rise in computational resources, have allowed for remarkable gains in the capabilities of ML algorithms, igniting a competitive landscape in this field. These algorithms, initially devised by humans, now actively participate in decision-making and policy formation for the same people who created them. The advantages of these algorithms are vast, as they enhance efficiency, accuracy, and speed in various domains. They contribute to improved legal outcomes [1], streamlined lending and hiring processes [2], and optimized allocation of resources and benefits [3]. Harnessing the power of ML to develop equitable and efficient systems can catalyze both social and economic progress.

Trustworthy ML. The initial belief that more data would result in better decision-making in the world of ML was quickly shattered as it became clear that accurate algorithms alone are not enough to make responsible decisions [4], [5], [6]. The significance of trustworthiness in ML can be explored by making an analogy with the stages of human development. Consider a child who inherits characteristics from their parents – this is akin to the initial model selection in ML, taking into account the mathematical limitations inherent in the chosen structure. As the child matures, they absorb crucial knowledge during their formative years - comparable to an ML model being trained with carefully selected datasets. The child's interaction with their socio-economic environment shapes their behavior and choices, just as an ML model's responses are influenced by the dataset it interacts with and the feedback it receives. The child experiences both opportunities and limitations in society, just as an ML model's functionality is affected by the boundaries of its mathematical design and the quality of its datasets. Ultimately, the child matures into an individual whose ethical decisions impact those around them, much like an ML model that must make responsible decisions affecting real-world outcomes. To develop a trustworthy ML pipeline, each element in the learning cycle, like each stage in a child's life, carries shared responsibility. In this research, we focus on understanding and integrating the two main pillars of trustworthy ML: Privacy and Fairness.

Privacy. Information privacy refers to an individual's right to maintain a certain level of control over how their personal data is gathered and utilized [7]. Take, for instance, a photo shared by an individual on social media platforms, intended solely for communication and social interaction. Even basic data mining techniques can extract sensitive information from this image, which could be exploited by malicious attackers. Aspects like ornaments, background, and facial features may inadvertently disclose the individual's religion, geographical location, gender, race, or other sensitive details. This raises the question of how much control users have over their own data. Expanding this concept to the vast quantities of data utilized in modern ML models highlights the importance of

S. Shaham, A. Hajisafi*, B. Krishnamachari, and C. Shahabi are with the University of Southern California, Los Angeles, CA 90089 USA (e-mail: sshaham@usc.edu; hajisafi@usc.edu; bkrishna@usc.edu; shahabi@usc.edu).

M. K. Quan* and P. N. Pathirana are with Deakin University, VIC, Australia (e-mail: m.quan@deakin.edu.au; pubudu.pathirana@deakin.edu.au).

D. C. Nguyen is with the University of Alabama in Huntsville, Huntsville, AL 35899 USA (e-mail: Dinh.Nguyen@uah.edu).

C. Peris is with Amazon, Cambridge, MA 02138 USA (e-mail: perisc@amazon.com).

G. Ghinita is with Hamad Bin Khalifa University, Doha, Qatar (e-mail: gghinita@hbku.edu.qa).

*Equal contribution.

privacy for ensuring trustworthy ML. Incidents like the 2020 Facebook scandal [8] and the Edward Snowden revelations [9] underscore the critical nature of user data privacy in the context of ML.

Fairness. From a different perspective, while privacy deals with the extent of control over data, fairness aims to ensure that the revealed user information is handled fairly and equitably. The philosophical notions of fairness have existed for centuries [10]; however, with the rapid growth of ML, algorithmic fairness and its application to ML have emerged as some of the most critical challenges of the decade. Unfortunately, models intended to intelligently avoid errors and biases in decision-making have themselves become sources of bias and discrimination within society. Various forms of unfairness in ML have raised concerns, including racial biases in criminal justice systems [11] and disparities in employment [12] and loan approval processes [13]. The entire life-cycle of an ML model – encompassing input data, modeling, evaluation, and feedback – is vulnerable to both external and inherent biases, leading to unjust outcomes. Compounding the issue is the tendency of the pipeline's life-cycle to amplify biases due to oversimplification and assumptions made throughout the process. Moreover, unlike the concept of privacy, for which there are well-defined and accepted metrics, the large number of varied and often conflicting definitions of fairness presents a significant challenge in establishing trustworthy ML systems.

Privacy vs Fairness. Investigations into privacy and fairness have often been carried out separately, without a holistic comprehension of how these two goals intertwine. Although the trade-offs between privacy and utility, as well as fairness and utility, are well-established, the complex relationship between these objectives remains less clear. Several studies such as [14] and [5], indicate the presence of trade-offs, while others, like [15] and [16], consider them to be in harmony. Given the lack of studies elucidating their interconnection, there is an urgent need for more research to uncover the link between these two goals, ultimately paving the way for truly responsible ML models.

Motivation. Pursuing privacy and fairness as separate objectives may appear intuitive, yet this approach is fraught with significant issues. First, engineers and researchers often find that the attainment of even one of these goals can significantly impact a model's performance, necessitating careful alignment of both objectives. Second, research exploring how the achievement of one goal influences the other remains limited. This knowledge gap introduces uncertainty concerning the model's reliability. As a result, our goal is to bridge this divide between privacy and fairness, traditionally pursued as independent objectives. We aim to establish a foundation for more advanced techniques facilitating their concurrent implementation, honoring these elements as the two primary pillars of trustworthy ML models.

Contribution. In this comprehensive survey, we present an in-depth examination of the main concepts in privacy and fairness by analyzing nearly 200 recent studies in the field. We explore these approaches across four primary aspects of ML, namely, Supervised Learning (SL), Unsupervised Learning (UL), Semi-Supervised Learning (SSL), and Reinforcement Learning (RL), with the aim of consolidating terminology and ideas. For instance, we collate and explain 15 distinct fairness notions to facilitate a better understanding of the principles. By establishing a solid comprehension of privacy and fairness across various ML techniques, we offer an extensive review of existing research on architectures designed to meet these goals, the interplay between the objectives, their concurrent implementation, and ultimately, their manifestation in several applications. Moreover, we identify several key unresolved questions and challenges in understanding two objective functions, from large language models to the disparate impact of privacy-preserving methods in ML.

The rest of this survey is organized as follows. A detailed examination of privacy within the realm of ML is presented in Section II. Concepts of fairness and algorithms to ensure it are then discussed in Section III. The intersection of privacy and fairness is the central focus of Section IV. Open issues and potential directions are explored in Section V. Conclusions are drawn in Section VI. A comprehensive map, illustrating the sections and subsections, can be found in Appendix A. To the best of our knowledge, this is the first survey that attempts to provide a critical review of privacy and fairness in ML. Some of the most relevant and recent surveys are reviewed in Table I.

II. Privacy

A. Preliminaries

ML has transformed industries like healthcare [28], transportation [29], and finance [30] with its predictive capabilities. However, the use of personal data in ML raises significant privacy concerns, including the potential misuse of sensitive information [31]. Techniques like differential privacy (DP) [32] and homomorphic encryption [33] are being developed to mitigate these risks. Continued research is essential to comprehensively address privacy concerns across various ML methodologies.

B. Privacy Techniques

1) Differential Privacy (DP)

DP is a privacy protection method commonly used in different stages of the ML pipeline to enhance privacy of individuals. In this section, we will examine the concepts and definitions, common DP mechanisms, and applications of DP in various ML techniques.

a) Notions and Definitions

Definition II.1 (ϵ-Differential Privacy[32])**.** A randomized algorithm $\mathcal{M}$ is said to be (ϵ, δ)-differentially private if, for any two datasets D_1 and D_2 that differ in only one data point, and any subset of the range of $\mathcal{M}$, the following holds:

$$Pr[M(D_1) \in S] \leq e^{\epsilon} Pr[M(D_2) \in S] + \delta, \qquad (1)$$

where ϵ and δ are privacy parameters, and S is any subset of the range of $\mathcal{M}$. This inequality ensures that the probability of observing a certain output of $\mathcal{M}$ on a dataset D_1 is almost the same as the probability of observing the same output on a dataset D_2 that differs in only one data point, with the exception of a small amount of random noise controlled by ϵ and δ. The parameter ϵ controls the strength of the privacy

Table I: Comparison of recent related works and our paper's key contributions on privacy and fairness in ML.

Paper	Key topic	ML categories				Key contributions		Highlights
		SL	UL	SSL	RL	Privacy	Fairness	
[17]	Privacy-preserving collaboration in ML	✓	✓	×	×	✓	×	Pioneering study concentrating on collaborative ML privacy needs and limitations
[18]	Limitations of DP in ML applications	✓	×	×	×	✓	×	DP assessment: flaws, trade-offs, ML implementation
[19]	Fairness-aware ML in different datasets	✓	✓	×	×	×	✓	Fairness in ML through in-depth real data analysis
[20]	Algorithmic fairness in ML	✓	×	×	✓	×	✓	Overview of identifying, measuring, and improving algorithmic fairness
[21]	Fairness in graph mining	✓	✓	✓	×	×	✓	Fairness in graph algorithms: measures, benchmarks, and research directions
[22]	Privacy-preserving ML	✓	✓	✓	×	✓	×	Identifying gaps and challenges in privacy preservation for ML
[23]	Privacy defense trade-offs in ML evaluation	✓	×	×	×	✓	×	Balancing privacy and utility in ML defense evaluation
[24]	Privacy-preserving ML	✓	×	✓	×	✓	×	Integration of privacy techniques in ML for data-driven applications
[25]	In-processing fairness mitigation	✓	✓	✓	×	×	✓	Categorization of explicit and implicit methods in achieving fairness
[6]	Fairness and bias in AI systems	×	×	✓	×	×	✓	Taxonomy of fairness definitions for mitigating biases in AI
[26]	Fair clustering	×	✓	×	×	×	✓	Organized overview with new insights and classifications in fair clustering
[27]	Privacy-Preserving DL in MLaaS	×	✓	✓	×	✓	×	Adversarial models, attacks, and solutions in privacy-preserving DL
Our paper	Privacy-Fairness Interrelation in ML	✓	✓	✓	✓	✓	✓	Thorough review of privacy, fairness in ML, examining impact, architectures, and research gaps

guarantee, with lower values providing stronger privacy protection, while δ is a parameter that accounts for the probability that the privacy guarantee is violated due to the randomness introduced by the algorithm.

Definition II.2 ($L1$-Sensitivity[34])**.** L1-sensitivity is a measure of how much the output of a function changes when a single data point is added or removed from a dataset. It is defined as the maximum absolute difference between the output of the function on two adjacent datasets that differ in only one data point. Formally, given a function $f : \mathcal{D} \rightarrow \mathbb{R}^n$ that maps datasets in domain $\mathcal{D}$ to vectors in $\mathbb{R}^n$, the L1-sensitivity of f is defined as:

$$\Delta f = \max_{d \in D, d' \sim d} ||f(d) - f(d')|\,|_1, \tag{2}$$

where d' is the neighboring dataset that differs from d by a single data point, and $||\cdot||_1$ denotes the L1-norm. Intuitively, L1-sensitivity captures the largest change that can occur in the output of f due to the presence or absence of a single data point. It is a fundamental parameter in DP, as it determines the amount of noise that needs to be added to the output of f to achieve a desired level of privacy protection.

Table II: Comparison of Laplace and Exponential Mechanisms in DP.

Mechanism	Description	Pros	Cons
Laplace	Adds independent noise drawn from a Laplace distribution to the true output, proportional to the sensitivity of the query and inversely proportional to the privacy budget.	Simple implementation, provides strong privacy guarantees, and works efficiently for simple queries.	Produces noisy results, calibration of noise parameter can be difficult, may not perform well for high-dimensional data or complex queries.
Exponential	Adds independent noise drawn from an exponential distribution to the true output, proportional to the sensitivity of the query and inversely proportional to the privacy budget.	More precise results than Laplace, can perform well for high-dimensional data or complex queries.	Requires more sophisticated implementation, may be vulnerable to adaptive attacks, calibration of noise parameter can be challenging.

b) Common Mechanisms in DP

DP approaches involve the addition of controlled noise to data to safeguard the privacy of people while preserving the accuracy of analytic results. The *Laplace mechanism* and *Exponential mechanism* are two often used differentially private mechanisms, which are detailed as follows.

Laplace Mechanism The Laplace mechanism [35] is a method for achieving DP by adding random noise to the output of a query in a way that satisfies DP guarantees. Specifically, given a function $f : D \to R$ that we want to compute on a dataset D, the Laplace mechanism adds random noise to $f(D)$ according to the following formula:

$$f(D) + Lap\left(\frac{\Delta f}{\epsilon}\right), \tag{3}$$

where $Lap(\Delta f/\epsilon)$ is a random variable drawn from the Laplace distribution with mean 0 and scale parameter $\Delta f/\epsilon$, where Δf is the sensitivity of the function f and ϵ is the privacy parameter that controls the amount of noise added. More formally, the Laplace distribution is defined as:

$$Lap(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right), \tag{4}$$

where μ is the mean and b is the scale parameter. In the case of the Laplace mechanism, the mean is 0 and the scale parameter is $\Delta f/\epsilon$, so the Laplace distribution becomes:

$$Lap(x \mid 0, \Delta f/\epsilon) = \frac{1}{2(\Delta f/\epsilon)} \exp\left(-\frac{|x|}{\Delta f/\epsilon}\right). \tag{5}$$

Adding Laplace noise to the output of $f(D)$ in this way ensures that the output is differentially private with parameter ϵ. The amount of noise added is proportional to the sensitivity of the function f, with higher sensitivities resulting in more noise being added to the output.

Although the Laplace mechanism is commonly used to achieve DP in ML, it has several limitations [36] [37]. The amount of noise added to the data depends on the sensitivity of the function being computed, which can be significant for some functions. This can lead to a considerable loss of output accuracy, making it difficult to obtain meaningful results. Moreover, the Laplace mechanism assumes that the data is continuous and unbounded, which may not always hold for all datasets. Furthermore, the Laplace distribution employed in the mechanism may not be optimal, as it assumes that the noise added to the data is symmetric, which may not be the case in reality. However, this technique is also commonly employed in terms of DP. For example, methods have been devised in [38] and [39] for releasing counts on specific types of data, such as time series. The authors from [40], [41] and [42] concentrate on releasing histograms, while other authors in [43], [44] present ways for reducing the worst-case error of a specified set of count queries.

Exponential Mechanism The exponential mechanism is proposed in [45], which is a privacy-preserving approach that selects an item from a dataset based on a specific objective function. It ensures the confidentiality of individuals in the dataset while maximizing the objective function. Formally, let $f : D \to R$ be a function that maps a dataset D to a real number. The exponential mechanism selects an output $d \in D$ with probability proportional to the exponential of the privacy loss incurred by releasing $f(d)$, scaled by a parameter $\epsilon > 0$, which controls the amount of privacy protection:

$$P(M(D) = d) \propto \exp\left(\frac{\epsilon f(d)}{2\Delta f}\right), \tag{6}$$

where ϵ is the privacy budget and $\propto$ denotes proportionality. The denominator $2\Delta f$ is used to scale the noise so that it is proportional to the sensitivity of the objective function.

In practice, the exponential mechanism is used when we want to select an element from a dataset that satisfies a certain property while minimizing the disclosure of information about the other elements in the dataset. For example, we might want to select a movie from a database that satisfies certain genre preferences while minimizing the disclosure of information about the users who rated the other movies in the database.

Popular way for creating DP in ML, the exponential mechanism has certain limitations[46]. When the dataset is huge, the exponential process can be computationally costly. In addition, as the privacy parameter falls, the precision of the result degrades. In addition, the exponential technique is only appropriate for functions with a low sensitivity value. If the function has a high sensitivity value, the mechanism will add an excessive amount of noise to the output, reducing

its precision. In certain instances, such as when the output space is discrete or the objective function is non-convex, the exponential mechanism might be biased. Despite these limitations, this permits DP solutions for a variety of intriguing issues with non-real outputs. As an illustration, the exponential mechanism has been used in the publication of audition results [45], coresets [47], support vector machines [48], and frequent patterns [49].

c) Spectrum of DP Variations

Differential Privacy Stochastic Gradient Descent (DP-SGD) DP-SGD emerged from the convergence of two important concepts in the field of ML: Stochastic Gradient Descent (SGD) and DP. SGD is an iterative method for optimizing an objective function and has been extensively used in ML, especially in the training of large-scale deep neural networks. The idea was to provide formal privacy guarantees when disclosing statistical information about a dataset. In 2016, Abadi et al. [50] successfully combined these concepts to develop DP-SGD, a variant of SGD that offers strong privacy guarantees by incorporating differential privacy into the optimization process.

The DP-SGD algorithm begins by sampling a minibatch from the dataset. For each instance in the minibatch, the gradient $\nabla L(\theta; x)$ of the loss function L with respect to the model parameters θ is computed. This results in a vector of gradients for the minibatch. The next crucial step in DP-SGD is gradient clipping. This process involves limiting the L_2 norm of each individual gradient vector to a predefined threshold C. In mathematical terms, this operation can be expressed as:

$$\nabla L_{\text{clipped}}(\theta; x) = \min\left(1, \frac{C}{\|\nabla L(\theta; x)\|}\right) \nabla L(\theta; x). \quad (7)$$

This gradient clipping operation ensures that the contribution of each individual instance to the gradient computation is limited, thereby mitigating the impact of outliers and reducing the sensitivity of the output to changes in the input data, a key requirement for DP. After gradient clipping, the algorithm computes the average of the clipped gradients and adds calibrated Gaussian noise to this average. If G represents the average of the clipped gradients, the noisy gradient G_{noisy} is given by:

$$G_{\text{noisy}} = G + \mathcal{N}(0, (\sigma C)^2 \mathbf{I}), \quad (8)$$

where $\mathcal{N}(0, (\sigma C)^2 \mathbf{I})$ represents multivariate Gaussian noise with mean 0 and covariance matrix $(\sigma C)^2 \mathbf{I}$, and $\mathbf{I}$ is the identity matrix. The model parameters θ are then updated using this noisy gradient.

DP-SGD is primarily used in scenarios where models need to be trained on sensitive data while preserving privacy. For instance, in healthcare, DP-SGD could be used to build predictive models using patient data without compromising individual privacy [51]. DP-SGD has also been used in federated learning [52], a paradigm where the model is trained across multiple decentralized edge devices, maintaining data on the original device. Several libraries and frameworks have been developed for implementing DP-SGD. Google's TensorFlow Privacy library provides a version of DP-SGD that can be used with TensorFlow models [53]. Another library is PyTorch-DP (now Opacus) [54], which provides an implementation for PyTorch models. These libraries provide convenient tools to add privacy-preserving capabilities to ML models with minimal code changes.

Differential Privacy for Support Vector Data Description (DP-SVDD) Support Vector Data Description (SVDD) [55] is a one-class classification method that is often used for anomaly detection. The main idea behind SVDD is to find a hypersphere in the feature space that encapsulates the majority of the data points. This hypersphere is described by its center and radius, and it is found by solving an optimization problem that aims to minimize the radius while penalizing data points that lie outside the hypersphere. Mathematically, the SVDD problem can be formulated as follows:

$$\begin{aligned} &\text{Minimize:} \quad R^2 + C\sum \xi_i, \\ &\text{Subject to:} \quad ||\phi(x_i) - a||^2 \leq R^2 + \xi_i \quad \text{and} \quad \xi_i \geq 0. \end{aligned} \quad (9)$$

In this formulation, R is the radius of the hypersphere, a is the center, $\phi(x_i)$ is the mapping of data point x_i in the feature space, ξ_i is the slack variable that allows data points to lie outside the hypersphere, and C is a regularization parameter that controls the trade-off between the volume of the hypersphere and the errors. DP-SVDD, first introduced in [56], is a method that combines the principles of SVDD with those of DP to create a privacy-preserving one-class classification model. The method involves two main phases. In the first phase, the goal is to train a SVDD model while ensuring differential privacy. The center of the hypersphere in the SVDD model is represented as a weighted sum of the mapped data points. To ensure DP, the center of the hypersphere is perturbed by adding noise. This noise is drawn from a Laplace distribution. The perturbed center, $\hat{a}$, is then given by:

$$\hat{a} = a + l = \sum_{i=1}^{n} b_i \phi(x_i) + l. \quad (10)$$

In this formulation, a is the center of the hypersphere, b_i are the dual variables, $\phi(x_i)$ is the mapping of data point x_i in the feature space, l is the Laplace noise. The sensitivity is a measure of how much the output of a function can change when a single data point is added or removed from the dataset. In the second phase, the input space is partitioned into separate regions using a dynamical system based on the differentially private support function from the first phase. This dynamical system is defined by the gradient of the support function:

$$\frac{dx}{dt} = \nabla \hat{f}(x), \quad (11)$$

where $\hat{f}(x)$ is the differentially private support function. Regions, associated with Equilibrium Points (EPs) of the dynamical system, are labeled using a noisy count of class labels of converging data points. The privacy-preserving predictions are released by publishing private EPs and labels. A new data point's label is predicted based on its region, determined by the EP it converges to. The privacy of predictions is ensured by

the differential privacy of the support function and the noisy count.

The preceding discussion regarding the DP has improved our understanding of the definitions and mechanisms of the DP, as well as the limitations of each mechanism, which are summarized in Table II. Additionally, a comprehensive examination of the real-world applications of DP in supervised, unsupervised, semi-supervised, and reinforcement learning is presented in Appendix B.

2) Homomorphic Encryption (HE)

Homomorphic Encryption (HE) represents a cryptographic technique that confers the capacity to perform computations on encrypted data without the need for decryption. Such an approach stands out as a promising means for preserving data privacy while enabling useful computations to be conducted on it. The following section reviews and categorizes the concepts, prevalent HE mechanisms, and a range of applications of HE in the context of ML techniques.

a) Notions and Definitions

Definition II.3 (Homomorphic Encryption[57])**.** Homomorphic encryption enables computations to be performed on ciphertexts without the need to decrypt them first. Mathematically, let f be an algebraic function and Enc and Dec be encryption and decryption functions respectively. Homomorphic encryption allows for the following equation to hold:

$$f(\mathrm{Dec}_k(\mathrm{Enc}_k(m_1)) \circ \mathrm{Dec}_k(\mathrm{Enc}_k(m_2))) = \mathrm{Enc}_k(f(m_1 \circ m_2)), \tag{12}$$

In this equation, m_1 and m_2 are plaintext messages that are encrypted under the same key k using HE. The function f is a homomorphic function that operates on the plaintext messages, and the operator $\circ$ represents the algebraic operation that f preserves. The equation shows that applying f to the plaintext messages m_1 and m_2 and then encrypting the result under the key k is equivalent to first encrypting the plaintext messages separately, applying the decryption function Dec_k to each ciphertext, performing the algebraic operation $\circ$ on the resulting plaintexts, and then encrypting the result again under the key k. This property allows computations to be performed on encrypted data without ever revealing the plaintext to the party performing the computation.

b) Typical Schemes in HE

There exist various schemes of homomorphic encryption, each possessing its unique merits and demerits. The *Fully Homomorphic Encryption*, *Partially Homomorphic Encryption*, and *Somewhat Homomorphic Encryption* are three frequently utilized HE schemes, explicated as follows.

Fully Homomorphic Encryption (FHE) FHE allows computations on encrypted data without decryption, leading to direct computation on ciphertexts and yielding an encrypted plaintext. It employs lattice-based cryptography [58] with ideal lattices to efficiently compute homomorphic operations. This type of cryptography is based on mathematical lattices, regular patterns of points. FHE carries out encryption and decryption on ideal lattices, which are sets of linear combinations of n independent vectors with integer coefficients in n-dimensional space. Mathematically, this can be expressed as:

$$L = a_1.v_1 + a_2.v_2 + ... + a_n.v_n | a_i \in Z, \tag{13}$$

where L is the lattice, v_1, v_2, ..., v_n are linearly independent vectors in n-dimensional space, and a_1, a_2, ..., a_n are integers. The security of FHE is based on the hardness of certain problems related to lattices, such as the Shortest Vector Problem (SVP) and the Closest Vector Problem (CVP) [59]. These problems are known to be difficult to solve in high dimensions, which provides the basis for the security of FHE. To perform homomorphic operations on ciphertexts in FHE, a technique called "bootstrapping" or "gating" is used. This technique involves decrypting the ciphertext using the secret key, performing a homomorphic operation on the resulting plaintext, and then encrypting the result using the public key. Mathematically, this can be represented as:

$$C' = Enc_{pk}(F(Dec_{sk}(C))), \tag{14}$$

where C is the original ciphertext, sk is the secret key, pk is the public key, F is the homomorphic operation, $Dec_{sk}(C)$ is the decrypted ciphertext, and Enc_{pk} is the encryption function using the public key.

FHE presents a spectrum of advantages and disadvantages [60]. On the one hand, FHE confers a paramount level of security as it allows for arbitrary computations on encrypted data without necessitating decryption. This characteristic proves exceptionally advantageous in domains such as ML and cloud computing, where privacy and security concerns are of utmost importance. Conversely, FHE exhibits a high level of computational complexity that can render it infeasible for certain applications [61]. Moreover, with each homomorphic operation, the size of the ciphertext augments, requiring extensive memory, which can become a significant impediment [62]. Despite these challenges, FHE remains an active area of research and development, with researchers continuously seeking ways to enhance its efficiency and transcend its limitations.

Partially Homomorphic Encryption (PHE) Partially Homomorphic Encryption (PHE) is a type of encryption scheme that enables computation on encrypted data without the need to decrypt it [63]. Mathematically, PHE is defined using algebraic structures such as groups, rings, or fields to enable certain types of computation, such as addition or multiplication, on ciphertexts while still maintaining the confidentiality of the underlying plaintext [64]. PHE schemes can be partially homomorphic, meaning that they support computations of only one type, such as addition or multiplication. For example, a PHE scheme that is homomorphic with respect to addition is defined by the following property:

$$Enc(m_1) + Enc(m_2) = Enc(m_1 + m_2), \tag{15}$$

where m_1 and m_2 are plaintext messages, Enc is the encryption function, and $+$ denotes addition in the plaintext space M. This property allows ciphertexts to be added together and then decrypted to obtain the sum of the corresponding plaintexts. Similarly, a PHE scheme that is partially homomorphic with

Table III: Comparison of Homomorphic Encryption Schemes.

Property	FHE	PHE	SHE
Supports addition	✓	✓	✓
Supports multiplication	✓	×	✓
Supports arbitrary circuits	✓	×	×
Computational complexity	High	Moderate	Low
Encryption/decryption speed	Slow	Moderate	Fast
Application examples	Cloud computing, privacy-preserving ML	Secure multi-party computation, secure function evaluation	Privacy-preserving data analysis, secure computation protocols

respect to multiplication is defined by the following property:

$$Enc(k \cdot m) = Enc(m)^k, \tag{16}$$

where k is a scalar value and m is a plaintext message. This property allows a ciphertext to be raised to a scalar power k without revealing the plaintext, but it does not allow multiplication of two ciphertexts to obtain a ciphertext that represents the multiplication of the corresponding plaintexts.
In reality, PHE is a powerful cryptographic technique that offers many benefits [65]. PHE allows for computations on encrypted data, enabling secure processing of sensitive data without its disclosure to unauthorized parties. Unlike FHE, PHE does not require heavy computational resources and is, therefore, much easier to implement in real-world applications. PHE can be implemented with relatively straightforward mathematical operations, making it both basic and effective. In addition, PHE can be used to build secure protocols for a range of applications, such as secure auctions, electronic voting, and secure multi-party computation. By keeping the data encrypted, PHE can ensure that sensitive information remains private while allowing authorized parties to perform meaningful computations on it. While there are some limitations to PHE [66], such as its limited computational capacity and susceptibility to attack if not implemented correctly, the benefits of this technique make it a valuable tool in a variety of situations.

Somewhat Homomorphic Encryption (SHE) Somewhat Homomorphic Encryption (SHE) is a form of encryption that permits certain calculations on encrypted data without revealing the original data [67]. Using a polynomial representation of the plaintext and encrypting it with a public key is a central concept of SHE [68]. By manipulating the coefficients of the polynomial, it is possible to perform computations on the ciphertext. A commonly used example of a SHE scheme is the BGV (Bajard, Gentry and Vaikuntanathan) scheme [69]. Consider the BGV scheme, which operates over the polynomial ring $R_q = \mathbb{Z}[x]/(xn+1)$, where q is a prime number that determines the security level and n is the degree of the polynomial. R_2 denotes the set of polynomials with coefficients in $\{0,1\}$, which is the plaintext space.
To encrypt a plaintext polynomial $m(x)$, the BGV scheme first generates a random polynomial $r(x)$ with coefficients in $\{0,1\}$. It then computes the ciphertext polynomial $c(x)$ as:

$$c(x) = r(x) * pk + m(x) * 2^k \pmod q, \tag{17}$$

where pk is the public key, k is a positive integer, and $*$ denotes polynomial multiplication. The random polynomial $r(x)$ serves as a noise term that hides the underlying plaintext, while the term $m(x)*2^k$ ensures that the ciphertext coefficients are sufficiently large to prevent decryption attacks. Conversely, to decrypt a ciphertext $c(x)$, one needs to compute:

$$m(x) = c(x) * sk \pmod q \bmod 2, \tag{18}$$

where sk is the secret key. The term $c(x) * sk$ cancels out the noise term $r(x)$, and yields the original plaintext polynomial $m(x)$. Finally, to perform a computation on two ciphertexts $c_1(x)$ and $c_2(x)$, one can simply add or multiply them as polynomials, and obtain the resulting ciphertext $c_3(x)$ as:

$$c_3(x) = c_1(x) + c_2(x) \text{ or } c_3(x) = c_1(x) * c_2(x). \tag{19}$$

Though SHE can enable computations to be performed on encrypted data without requiring the data to be decrypted, this scheme suffers from certain limitations that affect its practicality in certain scenarios [70]. These limitations stem from its lack of full homomorphic capabilities, which constrain the range of computations that can be performed. Furthermore, SHE is typically associated with higher computational overheads and requires more computational resources, which can affect its overall efficiency and practicality. Nevertheless, SHE offers several benefits [71], such as preserving the privacy of sensitive data, while allowing computations to be performed in a secure and confidential manner. This makes SHE a useful technique in scenarios where privacy and security are paramount, such as in the healthcare and financial sectors. Additionally, SHE can be used in conjunction with other cryptographic techniques, such as fully homomorphic encryption, to provide a more comprehensive security framework. Hence, despite its limitations, SHE remains a valuable and promising technique in the field of cryptography.

The preceding discourse on HE has enhanced our knowledge of its definitions, mechanisms, and the limitations of each scheme, as detailed in Table III. Additionally, a thorough examination of the practical applications of HE in supervised, unsupervised, semi-supervised, and reinforcement learning is provided in Appendix C.

3) Implementation of Privacy in ML

The implementation of privacy-preserving techniques in machine learning necessitates a judicious selection of algorithms tailored to the specific data characteristics and privacy requirements. For continuous data, such as releasing aggregate statistics while preserving individual privacy, the Laplace mechanism is often preferred, adding calibrated random noise drawn from a Laplace distribution to the true statistic. This perturbation, proportional to the query's sensitivity, hinders attackers from inferring individual data while maintaining reasonable accuracy. Conversely, for categorical data, the Exponential mechanism may be more suitable, sampling an output with probability proportional to the exponential of a utility function, balancing privacy and accuracy. Furthermore, implementing homomorphic encryption (HE) involves selecting an appropriate scheme (e.g., Brakerski/Fan-Vercauteren (BFV), Brakerski-Gentry-Vaikuntanathan (BGV), Cheon-Kim-Kim-Song (CKKS)), generating keys, encrypting data, performing computations on the ciphertext, and decrypting the result, enabling computations on encrypted data without compromising privacy.

III. Fairness

A. Preliminaries

The concept of fairness in society has been a recurring study subject throughout history [10]. Although early discussions were mainly philosophical, the rise of data and ML in the past decade has attracted tremendous attention to fairness in algorithms. As opposed to the initial perception of models and algorithms being trustworthy, soon it was realized that they could lead to severe unjust decisions, affecting especially individuals from disadvantaged groups. Perhaps, the most significant of such discoveries was revealed in an article published by ProPublica in 2016, highlighting the significance of algorithmic fairness. The article focuses on a software named COMPAS [72] designed to determine the risk of a person committing another crime and assist US judges in making release decisions. The investigation found that COMPAS was biased against African Americans as it had a higher rate of false positives for this group compared to Caucasians. This and numerous other examples indicate the necessity to quantify and mitigate unfairness-related issues in ML. In the remaining of this subsection, we discuss bias in ML, what the law says about the issue, and online tools available to address the problem.

1) Bias

The term "bias" in ML has a distinct meaning that is different from the typical understanding of the term in social and news contexts [73]. Bias is seen as the root cause of unfairness and is often tied to a specific term that indicates where in the process, the data is being distorted. Over time, many different types of bias have been introduced in the literature, some of which are subcategories of others, leading to confusion in properly defining each one. For interested readers, we have provided a thorough classification and visualization of bias in ML in Appendix D. In this categorization, we have grouped potential types of bias into four general categories: *A Biased World*, *Data Collection and Preparation*, *Model Training*, and finally, *Evaluation and Deployment*.

2) Philosophies of Fairness in Context

In the domain of work and employment, principles of fairness and non-discrimination guide the relationships among employees, employers, and labor unions. Two core fairness principles, often identified as 'Disparate Impact' and 'Disparate Treatment', are observed in this context. Disparate Treatment [74] acknowledges that unjust behaviors towards individuals due to their protected attributes, such as race, are unacceptable. An instance reflecting this principle in action could be prohibiting the exclusive skill examination of job applicants based on their ethnic group affiliation. Disparate impact [75] pertains to practices that inadvertently disadvantage a protected group, even though the policies implemented by organizations appear neutral on the surface. This principle recognizes that discrimination is not always direct, and it can affect individuals and groups in indirect ways. A classic example includes policies that, while appearing neutral, disproportionally impact members of a protected group in a negative manner [76].

In ML, fairness principles are implemented in various ways to uphold the aforementioned principles for sensitive attributes. Notably, different organizations provide guidelines on what constitutes sensitive attributes. The most commonly protected features include race, gender, religion, and national origin. For more detailed information, please refer to Appendix E, which provides a table of sensitive attributes identified by several organizations.

Structure In the following subsections, we thoroughly review fairness in supervised, unsupervised, semi-supervised, and RL. In each subsection, we start by defining fairness notions and definitions dedicated to the type of ML learner, followed by explaining the existing unfairness mitigation techniques for fair treatment of individuals and groups. The mitigation algorithms are divided into pre-processing, in-processing, and post-processing strategies. SSL lies at the intersection of supervised and unsupervised learning. To the best of our knowledge, there is no specific fairness notion proposed particularly for SSL, despite the existence of dedicated unfairness mitigation algorithms. Due to limited space, we have moved the discussions related to SSL to Appendix F.

B. Fairness in Supervised Learning

1) Notions and Definitions

As opposed to privacy, where at least for statistical databases, there is a consensus on DP, there does not exist such an agreement on a common notion for fairness. One suggested guideline is to select the notion based on the underlying application. This section reviews some of the most widely adopted fairness notions for supervised learning. In this context, we use terms notion and definition interchangeably. Also, deviation from a fairness notion is referred to as *discrimination level*. Discrimination is usually manifested as the absolute value of the difference in metrics for different groups. Moreover, we denote the set of sensitive attributes by A, all observed attributes by X, latent attributes not observed by U, true label to be predicted by Y, and finally, predictor by $\hat{Y}$.

We debut our discussions on notions with statistical parity, one of the primary group-level fairness notions.

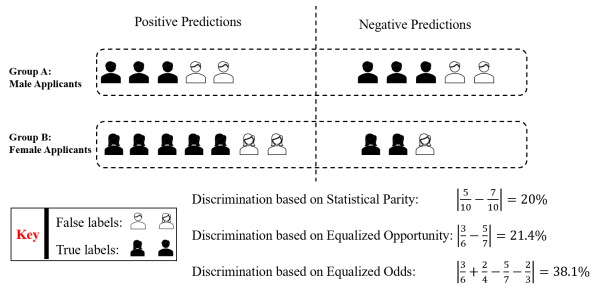


Figure 1: Computation of discrimination based on different metrics.

Definition III.1. (Demographic or Statistical Parity [4], [77]). A predictor $\hat{Y}$ satisfies demographic parity if:

$$P(\hat{Y}=1|A=0) = P(\hat{Y}=1|A=1). \tag{20}$$

Statistical parity dictates that regardless of an individual's group, they should have an equal chance of being assigned to a positive class. Figure 1 exemplifies statistical parity. Consider two groups of male and female job applicants and an ML model that decides whether a person should proceed for further evaluation in their application. Here, the likelihood of moving ahead with male and female applicants is $5/10$ and $7/10$, respectively. Hence, discrimination based on statistical parity is 20%. The notion of equalized odds, presented next, takes a step further and requires an equal true positive rate across groups.

Definition III.2. (Equalized Opportunity [4]). A predictor $\hat{Y}$ satisfies equal opportunity with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y,

$$P(\hat{Y}=1|A=0,Y=1) = P(\hat{Y}=1|A=1,Y=1). \tag{21}$$

That means the true positive rate should be the same for both groups. Going back to the example in Figure 1, the true positive rate for males and females is $3/6$ and $5/7$, leading to discrimination of 21.4% based on equalized opportunity. The next notion, equalized odds, dictates an even stricter fairness notion requiring equal true and false positive rates across groups.

Definition III.3. (Equalized Odds [4]). A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y if:

$$P(\hat{Y}=1 \mid A=0,Y=y) = P(\hat{Y}=1 \mid A=1,Y=y), \quad y \in \{0,1\}. \tag{22}$$

In the example, discrimination based on equalized odds is 38.1%. As can be seen, imposing higher fairness guarantees intuitively results in a higher percentage of discrimination.

Definition III.4 introduces the concept of calibration, which is a crucial idea borrowed from ML. This notion ensures that the confidence scores produced by the model can be interpreted as probabilities and is considered a group-level fairness notion.

Definition III.4. (Calibration [78], [79]). An ML model is said to be calibrated if it produces calibrated confidence scores. Formally, the outcome score R is said to be calibrated if for all the scores r in the support of R following stands,

$$P(y=1|R=r) = r. \tag{23}$$

Calibration ensures that the set of all instances assigned a score value r has an r fraction of positive instances among them. Note that the metric is defined on a group level, and it does not mean that an individual who has a score of r corresponds to r probability of a positive outcome. For example, given 10 people who are assigned a confidence score of 0.7, in a well-calibrated model, we expect to have 7 individuals with positive labels among them.

So far, the fairness definitions discussed were all focused on group-level fairness. In the following, two of the common notions to achieve fairness at an individual level are presented.

Definition III.5. (Counterfactual Fairness [80]). Given a causal model (U, V, F), where U, V, and F represent the set of latent (unobserved) background variables, the set of observable variables, and a set of functions defining the mapping $U \cup V \rightarrow V$, respectively, a predictor $\hat{Y}$ is considered counterfactually fair if, under any context $X=x$ and $A=a$, the following equation holds:

$$P(\hat{Y}_{A\leftarrow a}(U)=y|X=x,A=a) = P(\hat{Y}_{A\leftarrow a'}(U)=y|X=x,A=a). \tag{24}$$

This holds for all y and for any value a' attainable by A. Here, A, X, and $\hat{Y}$ represent the set of sensitive attributes, remaining attributes, and decision output, respectively. In other words, the model's predictions for a person should not change in a counterfactual world in which the person's sensitive features are different.

Definition III.6. (Individual Fairness by Dwork et al. [4]). For a mechanism $\mathcal{M}$ mapping $\boldsymbol{u}$ in the input space $\mathcal{X}$ to value y in the output space $\mathcal{Y}$, individual fairness is satisfied when for any $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{X}$:

$$d_{\mathcal{X}}(\boldsymbol{u},\boldsymbol{v}) \geq d_{\mathcal{Y}}(\mathcal{M}(\boldsymbol{u}),\mathcal{M}(\boldsymbol{v})), \tag{25}$$

where $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ and $d_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$.

To illustrate, let's consider the scenario of classification, where the classifier's predictor $\hat{Y}$ serves as the mapping mechanism. In the context of individual fairness, the fundamental idea is that two individuals who are alike in relevant ways should receive comparable outcomes. To operationalize this concept, we rely on two crucial distance metrics: (1) a similarity distance metric $d_{\mathcal{X}}$ that gauges how similar two individuals are to each other, and (2) a distance metric $d_{\mathcal{Y}}$ that quantifies the disparity between the distributions of outcomes.

2) Unfairness Mitigation Algorithms

a) Pre-processing Strategies

Reweighting. This approach focuses on modifying the significance of data points manifested as weights during the training to mitigate bias. The method employed in [81] estimates the probability of an individual from a group receiving a specific result and employs the ratios of these probabilities as reweighting factors during optimization. Nevertheless, in some instances, access to sensitive information may be restricted.

To tackle this issue, Lahoti et al. [82] learn the reweighting factor using adversarial learning. Furthermore, Roh et al. [83] suggest a two-tier optimization method that chooses specific mini-batch sizes to attain group fairness.

Representation learning. More recently, a thread of research has focused on changing data representation to improve fairness in classifiers. Several techniques have been proposed predicated on either philosophy of fairness through unawareness or fairness through awareness. In the former, the logic is to learn a new representation for individuals independent of their protected attribute while preserving other information, and the latter focuses on achieving fairness while considering protected group information. In [84], the authors modified dataset features to have similar distributions for both protected and unprotected groups, making it hard to distinguish between them. Zemel et al. [85] proposed mapping individuals to a new distribution that protects the protected group information while retaining other information. The authors in [86] follow up this idea by using a variational auto-encoder to make the sensitive attributes independent of the latent representation and applying further learning to that representation.

Label alteration. Label alteration aims to make classifiers fairer by adjusting the labels of training samples. Some works modify the labels to achieve an equal proportion of positive examples across protected groups [81], while others flip the labels of instances that have been determined to be discriminatory based on differences in treatment among similar samples [87].

b) In-processing Strategies

Regularizers and constraints. This strategy aims to add penalty terms to the classifier's objective function to either minimize the impact of sensitive features on prediction [88], or achieve similar False Positive and False Negative rates across populations [89], [90]. In a similar approach, Kamiran et al. [91] propose modifying the splitting criterion in decision trees to minimize the impact of sensitive features and maximize the information gain between the split feature and class label. The authors in several studies aim to enforce fairness notions constraints in optimization. This approach for statistical parity is discussed in [92], equalized odds and opportunity in [93], [94]. Quadrianto et al. [95] suggest using privileged learning to ensure fairness where sensitive features are only available during training.

Adversarial learning. The main concept within this group involves utilizing Generative Adversarial Networks (GANs) to optimize the effectiveness of a predictor while reducing its capability to forecast sensitive characteristics [96]. This approach can be implemented across various gradient-based learning models, such as classification and regression assignments.

Reweighting. The reweighting approach proposed part of pre-processing strategies has also been employed during the training. The proposed approach by Krasanakis et al. [97] trains an unweighted classifier, then learns weights for each sample and retrains the classifier to improve the fairness-accuracy trade-off. The iterative approach of improving reweighting factors helps to derive more accurate reweighting factors.

c) Post-processing Strategies

Transformation. This technique aims to modify the output scores to achieve higher fairness levels. One of the primary techniques in this category is Platt Scaling focused on improving miscalibration in ML models [98]. Calibration is improved by fitting the output scores to a logistic regression model. Histogram Binning and Isotonic Regression [99] also enhance calibration by fitting output scores to a monotonic function. To achieve individual fairness, the authors in [100] propose using c-fair polynomials, which map classifier scores to a polynomial and restrict scores of each individual by their sensitive feature distance. Petersen et al. [101] improve individual fairness by smoothing output scores using a similarity graph and Laplacian regularization. Kim et al. [102] present a Multiaccuracy Boost framework that improves accuracy across all subgroups, using iterations and weights to enhance predictions conducted by the auditor.

Thresholding. The proposed techniques in this category aim to adjust the label generation threshold of classifiers to make non-discriminatory decisions [103]. For instance, different threshold values are selected for protected groups in [104] to maximize accuracy while achieving statistical parity. Hardt et al. [105] optimizes threshold selection for each sensitive group for high utility and improved fairness. Similarly, the authors in [106] infer group-specific thresholds for a trade-off between accuracy and fairness. Lohia et al. [107] develop a bias-mitigation technique by targeting samples that inhibit individual bias from improving individual and group-level fairness notions.

C. Fairness in Unsupervised Learning

When data labels are unavailable, unsupervised ML algorithms are commonly used as opposed to supervised algorithms. However, evaluating fairness is more challenging in the absence of labels because there is no ground truth available for assessment. To address this issue, we will begin our discussion by examining individual and group-level fairness notions that have been suggested for unsupervised learning, followed by an exploration of mitigation algorithms.

1) Notions and Definitions

The majority of fairness concepts for unsupervised learning are based on the disparate impact doctrine, which seeks to achieve a comparable proportion of protected groups across all clusters. To begin our conversations on fairness concepts, we will start by examining the Balance metric, regarded as one of the fundamental definitions of group-level fairness in unsupervised learning.

Definition III.7. (Balance[108], [109]). Define the ratio of the protected group $b \in [m]$ in the entire dataset as r_b, and let $r_{a,b}$ represent this proportion in the generated cluster $a \in [k]$. The balance metric evaluates the disparity between these two ratios by defining $R_{a,b} = r_b / r_{a,b}$ and introducing the balance fairness concept as follows:

$$\min_{a \in [k], b \in [m]} \min R_{a,b}, 1/R_{a,b}. \tag{26}$$

The Balance metric produces values within the range of 0 to 1, where higher scores indicate a higher level of fair-

ness. This measure takes into account both the percentage of protected group members in the entire dataset and within individual clusters, with fairness achieved when the ratio remains consistent across all clusters. For example, consider a loan application clustering task where the protected attribute is race. If the overall dataset comprises 40% Black applicants, balance fairness necessitates that each cluster should also have approximately 40% Black applicants. This ensures that loan opportunities are not disproportionately allocated based on race.

Definition III.8. (Bounded Representation[110]). Let $r_{a,b}$ denote the ratio of protected group $b \in [m]$ in cluster $a \in [k]$. The (α, β)-bounded representations dictates that:

$$\beta \leq r_{a,b} \leq \alpha. \tag{27}$$

Bounded representation allows some degree of deviation in the proportion of protected groups within clusters. When the bounds are equal, it suggests that the proportion of protected groups in each cluster should be consistent with the overall ratio in the dataset. To illustrate, in clustering patients for medical treatment plans, where gender is the protected attribute, (α, β)-bounded representation with $\alpha = 0.6$ and $\beta = 0.4$ ensures that each cluster has at least 40% and at most 60% of patients from any specific gender. This prevents clusters from being overly skewed towards one gender, potentially leading to biased treatment recommendations.

Definition III.9. (Max Fairness Cost (MFC) [111]). Let I_b denote the ideal proportion of protected group $b \in [m]$ in clusters. Once the ideal ratio parameter is passed as input, the MFC notion is defined as

$$\max_{a \in [k]} \sum_{b \in [m]} |r_{a,b} - I_b|, \tag{28}$$

where $r_{a,b}$ denotes the ratio of protected group b in cluster $a \in [k]$.

Intuitively, MFC calculates the summation of all deviations from the ideal ratios for each protected group, and returns the maximum value. A lower MFC value indicates a higher degree of fairness. Setting the value of I_b equal to the ratio of the protected group in the original dataset (r_b) ensures that this ratio remains consistent across all clusters. This can be exemplified that in clustering individuals for social network analysis, where age is the protected attribute, MFC with an ideal proportion of 25% for each age group (young, middle-aged, and elderly) across all clusters would ensure age diversity within each cluster. A lower MFC value indicates a higher degree of fairness in age distribution across clusters.

Definition III.10. (Social Fairness [112]). Let C denote the cluster centers in k-means algorithm and $L(C, D_b)$ denote the k-means clustering cost, where D_b is the input error on the samples of the protected group $b \in [m]$. The social fairness notion is then defined as:

$$\max_{b \in [m]} \frac{L(C, D_b)}{|D_b|}, \tag{29}$$

Social fairness focuses on the maximum imposed loss on protected groups. For example, in clustering housing data for urban planning, where income level is the protected attribute, social fairness ensures that no income group experiences significantly higher clustering cost, preventing disparities in resource allocation or service provision based on income. Next, we focus on fairness notions proposed on the individual-level.

Definition III.11. (Fuzzy Individual Fairness [4]). For every two points x and y, and their respective distributions X and Y over clusters in a given fuzzy clustering algorithm, let $F(x, y)$ measure the similarity between the two datapoints, and let $D_f(X||Y)$ denote the statistical distance between their distributions. Fuzzy individual fairness requires the satisfaction of the following constraint:

$$D_f(X||Y) \leq F(x, y). \tag{30}$$

This notion aims to apply the individual fairness notion in [4] for fuzzy clustering. Common choices for measuring the statistical distance include the variations of f-divergence metric, such as KL-divergence, reverse KL-divergence, and the total variation distance. For example, consider clustering individuals based on their movie preferences, where similarity $F(x, y)$ is measured by the correlation between their movie ratings. If two individuals, Alice and Bob, have highly similar movie tastes (e.g., $F(x, y) = 0.9$), fuzzy individual fairness requires that their distributions over movie clusters should also be similar (e.g., $D_f(X||Y) \leq 0.9$). This ensures that Alice and Bob, with similar preferences, are not assigned vastly different probabilities of belonging to various movie clusters.

Definition III.12. (Individual Fairness [113]). This notion requires the average distance of every sample point to members in its own cluster to be smaller than its average distance to members of any other cluster. Formally, for a disjoint clustering of data denoted by $\mathcal{C} = C_1, , C_2, , \ldots, C_k$, and a distance metric d, for every sample point $x \in C_i$, the following inequality should hold:

$$\frac{1}{|C_i| - 1} \sum_{y \in C_i} d(x, y) \leq \frac{1}{|C_j|} \sum_{y \in C_j} d(x, y), , , , i \neq j. \tag{31}$$

The intuition behind the above notion is to ensure every sample point is associated with a cluster that has the highest average similarity. This can be illustrated by considering clustering researchers based on their research interests, where distance is measured by the difference in their publication topics. Individual fairness dictates that a researcher's average distance to other researchers in their assigned cluster (e.g., "machine learning") should be smaller than their average distance to researchers in any other cluster (e.g., "bioinformatics"). This ensures that researchers are clustered with those who share the most similar research interests, promoting collaboration and knowledge sharing within clusters.

2) Unfairness Mitigation Algorithms

a) Pre-processing Strategies

Fairlet Decomposition. The concept of fairlets, which was introduced in [114], aims to improve the fairness of

various clustering algorithms based on the Balance metric. The strategy involves dividing the data points into small groups, or so-called fairlets, before performing clustering in such a way that the disparate impact doctrine is maintained. Each fairlet is then represented by a single point, and a vanilla clustering algorithm is applied to these representative points. Because the representative points are reasonably fair, the final clustering also tends to be fair. In [108], near-linear algorithms are proposed for fairlet decomposition. Ahmadian et al. [115] employ the idea of fairlets for hierarchical clustering considering several objective functions such as revenue, value and cost.

Data Augmentation. Inspired by the approach in [116], Chhabra et al. [117] propose data augmentation as an efficient method for fair clustering. The method involves augmenting the dataset using a small subset of data to achieve a fairer clustering output after applying the algorithm. The authors propose a general bi-level formulation to address two problem settings: 1) using convex group-level fairness notions and convex center-based clustering objectives, and 2) using general group-level fairness notions and general center-based clustering objectives.

b) In-processing Strategies

Regularizers and Constraints. Built on the Balance metric, the authors in [118] incorporate fairness constraints in spectral clustering as well as providing empirical evidence that it is possible to achieve higher demographic proportionality at minimal additional cost in the clustering objective. Li et al. [119] introduce a fairness-adversarial term encouraging soft assignments that remain constant across various protected subgroups, resulting in a model that is not influenced by sensitive attributes. Zhang et al. [120] propose an approach for fairness in deep clustering. A regularization term based on the Balance notion is proposed and is combined with the clustering objective. Chai et al [121] propose to use Sinkhorn divergence to reduce differences in predicted soft labels among various demographic groups and to develop representations that are conducive to clustering. The requirement of equalized confidence is modeled as a regularization term during training using Sinkhorn divergence, with several additional regularizers for ensuring accuracy.

Alternating Objective. This approach focuses on entirely alternating between fairness and clustering objectives during unsupervised learning. Liu et al. [122] formulate the cost of clustering and fairness as a bi-objective optimization problem to achieve balance. Their approach is based on mini-batch k-means clustering, where the algorithm performs clustering based on the k-means during mini-batch updates. However, the algorithm also includes a series of swap-based steps to enhance the balance in clusters. The routine involves exchanging data points between the least balanced and well-balanced clusters after the mini-batch update. The method described in [123] combines the Kuulback-Leibler fairness term with the objective of center-based and graph-based algorithms. The approach involves conducting a separate update for the assignment of datapoints based on the fairness objective and clustering objective. The heuristic algorithm proposed in [124] focuses on achieving the proportionality fairness notion by alternating between fairness and clustering objective. The proposed algorithm achieves a $(1+\sqrt{2})$-proportional solution.

c) Post-processing Strategies

The majority of the methods proposed for the post-processing stage involve using linear programming to reassign data points according to fairness metrics. This approach is referred to as the LP formulation. Additionally, Simoes et al. [125] have recently explored an alternative approach based on *Data Perturbation* for fair clustering. The core concept of this approach is to perturb the assignment of data points to clusters in several iterations based on the "Rawls' difference principle" [126].

LP formulation. Considering disparate impact doctrine, the authors in [109] show that for a given clustering with l_p-norm objective including center-based approaches such as k-means and k-medoids, it is possible to have fair algorithms with a slight sacrifice in fairness constraint. In more detail, given any ρ-approximation algorithm for a given clustering objective, a $(\rho + 2)$-approximation solution exist for the best clustering, which satisfies fairness constraints. The objective is achieved by formulating and solving an LP optimization problem for fair assignment after clustering. Approaches in [110] and [127] also use alternative LP formulations for fair assignment of datapoints to centers. Esmaeili et al. [128] extend the approach to a scenario where data points are probabilistically assigned to groups instead of having a priori information on the group assignment. In [129], first the center-based clustering algorithm is applied to maximize the clustering objective. Then, using an LP formulation, clustering is improved considering the fairness objective. This is done by searching for the cluster with maximum violation of the fairness objective considering the upper bound required for the clustering objective and rounding the possibly fractional solution to a feasible integer solution using a network flow algorithm.

D. Fairness in Reinforcement Learning

RL is concerned with learning how to make decisions in an environment by maximizing some cumulative reward (or equivalently, minimizing some regret) through interacting with the environment and receiving feedback. The decisions made by RL agents can have a significant impact on individuals and society, making it essential to ensure that these decisions are unbiased and fair. Compared to other methods in which only the immediate impact of the decision-making algorithm is studied to mitigate unfairness, algorithms for fair RL aim to account for the long-term consequences of the agent's actions to ensure that they are unbiased [130].

In the context of fairness in RL, a significant focus is dedicated to addressing unfairness across various variants of the bandit problem. Bandit problems involve an agent repeatedly selecting from a set of arms, each associated with unknown reward distributions. The agent's objective is to determine an optimal *policy* that maximizes cumulative reward while receiving limited feedback on unchosen arms [131]. Bandit scenarios provide a tractable framework for exploring and developing fair decision-making algorithms in RL.

1) Notions and Definitions

In recent years, several perspectives have emerged for evaluating and improving the fairness of RL algorithms. Specifically, researchers have proposed and evaluated notions of fairness in RL from three distinct perspectives: *Meritocratic Fairness*, *Individual Fairness*, and *Proportional Fairness*. These perspectives can be quantified in different ways, depending on the specific goals and objectives of the RL algorithm.

Meritocratic Fairness. This perspective requires avoiding favoring less qualified individuals over more qualified ones. For example, in the context of bandits, this fairness notion indicates that it is unfair to preferentially select an arm with a lower expected reward over other available arms with higher expected rewards [132]. This ensures that the rewards are allocated fairly based on the arms' abilities. The following definitions are some examples of evaluating fairness from this aspect in literature.

Definition III.13. (δ-Fairness in Classic Bandits [132]). An algorithm $\mathcal{A}$ is deemed δ-fair if, with a probability of at least $1-\delta$ over history h, for all distributions $\mathcal{D}_1, ..., \mathcal{D}_k$, every $t \in [T]$, and all $j, j' \in [k]$:

$$\pi^t_{j|h} > \pi^t_{j'|h} \text{ only if } \mu_j > \mu'_j \tag{32}$$

Here, T represents a known horizon, $[k] = 1, ..., k$ denotes the set of arms, and $\mathcal{D}_1, ..., \mathcal{D}k$ are the unknown reward distributions of arms. μ_i is the unknown average reward of the i-th arm, and $\pi^t_{i|h}$ is the probability that $\mathcal{A}$ selects arm i given history h.

In essence, this definition suggests that selecting one arm over another is considered unfair if there is sufficient confidence to indicate that the chosen arm has a lower expected reward compared to the unselected one.

Definition III.14. (δ-Fairness in Contextual Bandits [132]). An algorithm $\mathcal{A}$ is considered δ-fair if, with a probability of at least $1-\delta$ over history h, for all sequences of contexts $x^1, ..., x^t$, all payoff distributions $\mathcal{D}^t_1, ..., \mathcal{D}^t_k$, every round $t \in [T]$, and all pairs of arms $j, j' \in [k]$:

$$\pi^t_{j|h} > \pi^t_{j'|h} \text{ only if } f_j(x^t_j) > f_{j'}(x^t_{j'}), \tag{33}$$

where $f_i : x_i \to [0, 1]$ denotes an unknown mapping from the context to the reward for each arm.

Individual Fairness. Individual fairness mandates that similar individuals should be treated similarly [4]. In the context of bandits, it means that arms with similar qualities should be selected by the algorithm with similar probability [133]. This ensures that no particular arm is consistently preferred over others that have similar expected rewards. The following definition provides a notion of individual fairness in the context of bandits.

Definition III.15. "(Smooth Fairness [133]). For a divergence function D, let $D(\pi_t(i) \| \pi_t(j))$ denote the divergence between Bernoulli distributions with parameters $\pi_t(i)$ and $\pi_t(j)$, and let $D(r_i \parallel r_j)$ denote the divergence between the reward distributions of the i-th and j-th arms. An algorithm $\mathcal{A}$ is $(\epsilon_1, \epsilon_2, \delta)$-fair with respect to the divergence function D if $\epsilon_1, \epsilon_2 \geq 0$, and $0 \leq \delta \leq 1$. With a probability of at least $1-\delta$ in every round t, for every pair of arms i and j, the following inequality should hold:

$$D(\pi_t(i) \parallel \pi_t(j)) \leq \epsilon_1 D(r_i \parallel r_j) + \epsilon_2. \tag{34}$$

In other words, if two arms have comparable reward distributions, a fair decision rule should treat them similarly by assigning them similar selection probabilities.

Proportional Fairness. Proportional Fairness in RL aims to ensure that each user, or in the case of bandits, each arm, is allocated a minimum guaranteed share of resources or pulls over time [134], [135]. By doing so, it guarantees that each arm is played at least a certain proportion of the time, thus ensuring a minimum level of exploration for all arms and preventing any particular arm from being unfairly favored over others. We present the following fairness notion as a way to quantify Proportional Fairness:

Definition III.16. (Asymptotic Fairness [134]). Let $d(t) = (d_1(t), \ldots, d_N(t))$ be a vector indicating whether each of the N arms is pulled at round t, where $d_i(t) = 1$ if arm i is played and $d_i(t) = 0$ otherwise. Moreover, let $r_i \in (0, 1)$ denote the required minimum fraction of rounds in which arm i is played. Algorithm $\mathcal{A}$ is called asymptotically fair if:

$$\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[d_i(t)] \geq r_i, \forall i \in [N]. \tag{35}$$

This means that as the number of rounds T approaches infinity, the expected fraction of rounds in which each arm is played should be equal to or greater than a prespecified fraction that is considered fair.

2) Unfairness Mitigation Algorithms

Owing to the inherent characteristics of RL methods, most techniques in the ML category are in-processs. We have classified the mitigation algorithms into three groups: Reward modification, action constraint, and algorithmic adjustments.

Reward modification. Methods in this category involve modifying the notion of regret or rewards that the agent receives during training to encourage fairness. For instance, in [136], researchers propose an extension to the conventional notion of regret, called *r-regret*, which incorporates fairness constraints to ensure that each arm is selected at least a prespecified fraction of the time at each time step in stochastic multi-armed bandit problems. In a separate study on Interactive Recommender Systems, a novel RL-based framework named FairRec is introduced to combine accuracy and fairness in the rewards function for making recommendations [137]. FairRec dynamically balances accuracy and fairness by incorporating user preferences and system fairness status into its state representations, which helps to improve fairness while preserving recommendation quality over time.

Action constraint. This category encompasses techniques that either constrain an agent's actions or adjust the action selection strategy during training to promote fairness. Joseph et al. [132] introduce *FAIRBANDITS* for achieving δ-fairness in stochastic bandits by modifying the UCB algorithm. When confidence intervals overlap, they recommend playing corresponding arms with equal probability. They also present a method for fairness in contextual bandit problems by converting the KWIK algorithm to a δ-fair contextual bandit algorithm

and vice versa. Jabbari et al. [138] extend δ-fairness to MDPs, ensuring no action is favored if it results in lower long-term discounted rewards. Their Fair-E^3 algorithm achieves fairness based on an approximate definition. Liu et al. [133] propose the notion of *smooth fairness*, a constraint based on the reward distributions as described earlier, and *fairness regret*, to measure calibration deviations. They show how to address these constraints in Bernoulli and Dueling bandit settings.

Algorithmic Modifications. This category involves modifying RL algorithms themselves to promote fairness. Some research in this area aims to guarantee a minimum number of times each arm is chosen in multi-armed bandit problems. For example, Chen et al. [139] incorporate constraints to ensure a minimum selection rate for each arm in contextual multi-armed bandit problems, suggesting an algorithm that minimizes regret for multiple contexts while maintaining fairness. Other studies target group-level fairness in RL-based decision-making. Huang et al. [140] frame personalized recommendations as a modified contextual bandit problem, introducing a fair algorithm called *Fair-LinUCB* to maintain parity in the expected mean reward of both the protected and unprotected groups. Wen et al. [141] propose fair sequential decision-making algorithms in MDPs that enforce fairness constraints based on average outcome quality for different subpopulations, aiming for demographic parity and equalized opportunity.

A key challenge in individual fairness is quantifying individual similarity. While many studies assume such a metric, Gillen et al. [142] propose learning a similarity metric during decision-making in contextual bandits setting, relying on an oracle that can identify fairness violations without explicitly providing a quantitative metric. Ge et al. [143] introduce *MoFIR*, a framework that balances fairness and utility in recommendation systems. The authors use Multi-Objective Reinforcement Learning to learn an optimal recommendation policy. MoFIR extends the Deep Deterministic Policy Gradient algorithm by incorporating a conditioned network that considers decision-maker preferences and outputs Q-value vectors. This approach aims to address fairness concerns while maximizing the effectiveness of recommendations.

E. Implementing Faireness in ML

Implementing fairness in ML requires a systematic approach involving careful selection of fairness metrics tailored to the application's ethical and legal context, such as equalized opportunity for ensuring equitable outcomes across demographics. Thorough data bias assessment is crucial to identify disparities in representation or labeling that could lead to unfair treatment. Algorithm selection depends on the specific needs and data characteristics, with options like reweighing for addressing biased labels and adversarial debiasing for mitigating biased features. Rigorous performance evaluation using metrics like accuracy, fairness (assessed with the chosen metric), and model explainability tools such as SHAP and LIME is essential to ensure fair and unbiased outcomes.

IV. Privacy & Fairness

A. Architectures

In this section, we introduce five prominent architectures specifically developed to tackle the dual challenges of privacy and fairness in ML. The visual representations of these architectures are presented in Figure 2. The first approach, Architecture A, ensures privacy and fairness by initially applying privacy-preserving algorithms to data, creating a sanitized dataset for fair training using techniques like the privacy-preserving k-means algorithm [144] and fairlet decomposition [114]. Architecture B achieves simultaneous privacy and fairness during model training by directly applying fair and private learning methods, such as those in [5] and [145]. Architecture C sanitizes sensitive user attributes with privacy-preserving methods before processing all attributes through the ML pipeline, applying fairness techniques at various stages as in [146]. Architecture D leverages Federated Learning (FL) for decentralized model training while maintaining data privacy, using Secure Aggregation [147] and addressing fairness through local and global strategies, exemplified in [148]. Finally, Architecture E involves privacy-preserving fairness auditing with Secure Multiparty Computation (MPC), enabling collaboration between a company and an auditing authority without exposing private information, as discussed in [149].

B. Impact of Privacy on Fairness

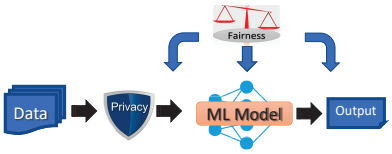
The aim of this subsection is to comprehend the ways in which privacy-preserving methods affect the fair treatment of individuals and groups. The subsequent discussion reviews two perspectives on perspectives on the potential positive and negative impacts of privacy on fairness: one focused on ways in which they are aligned, and another focusing on ways in they contrast with each other.

1) Aligned

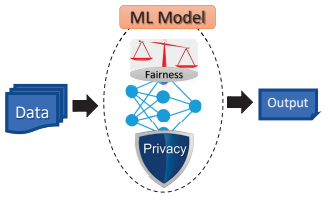
To begin our discussion, we examine publications that argue for the compatibility of privacy and fairness objectives. Pannekoek et al. [16] used a neural network with three fully connected layers, employing "reject option classification" for fairness and the DP version of the Adam optimizer [53] for privacy. Their model showed superior fairness and maintained high accuracy without performance decline as privacy measures increased. Khalili et al. [15] demonstrated that the exponential mechanism for DP can also achieve fairness when used as a post-processing approach. Sarhan et al. [150] explored the effect of DP on fairness in FL, finding that DP reduces discrimination in both local and global DP scenarios but requires parameter tuning due to a trade-off between privacy and fairness. Lyu et al. [151] proposed adding a noise layer to text features to achieve DP before classification, generally decreasing bias. Maheshwari et al. [152] integrated DP and adversarial learning, focusing on Equalized Fairness in NLP, and demonstrated that privacy and fairness are mutually supportive by perturbing text encoder outputs to achieve DP and employing a classifier with an adversarial branch to foster fairness.

2) Contrasting

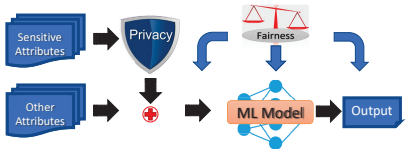
On the contrary, some studies argue that a trade-off exists between privacy and fairness. Sanyal et al. [153] explored the



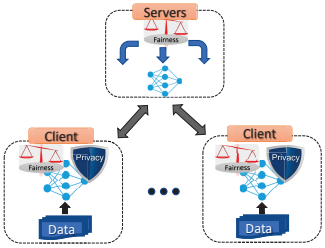
(a) Architecture A: Ensures privacy and fairness by sanitizing data before training.



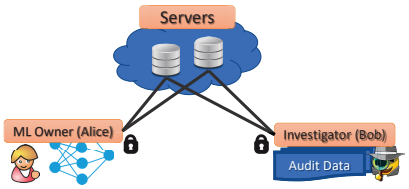
(b) Architecture B: Achieves privacy and fairness simultaneously during model training.



(c) Architecture C: Sanitizes sensitive attributes before processing through the ML pipeline.



(d) Architecture D: Uses Federated Learning for decentralized training while maintaining privacy.



(e) Architecture E: Employs Secure Multiparty Computation for privacy-preserving fairness auditing.

Figure 2: Architectures Enabling Simultaneous Implementation of Privacy and Fairness in ML.

use of Differential Privacy (DP) in classifiers, emphasizing accuracy discrepancy as a fairness metric. Their analysis revealed challenges in maintaining high accuracy for minority groups while ensuring strict privacy, particularly in datasets with a long-tailed distribution. Relaxing accuracy requirements was shown to enhance privacy and fairness. Bagdasaryan et al. [14] investigated neural networks trained with DP-SGD, finding that the privacy mechanism amplified bias towards more represented classes, affecting tasks like gender classification and sentiment analysis. Du et al. [154] examined DP's impact on outlier detection, highlighting how added noise can obscure rare training examples, reducing the model's ability to detect outliers effectively. Chen [5] discussed trade-offs in semi-private settings where a fraction of sensitive data is clean while privacy measures protect the rest, illustrating complexities in balancing privacy and fairness objectives.

C. Impact of Fairness on Privacy

There is a notable gap in understanding how algorithmic fairness impacts privacy, despite its significant implications. Strategies aimed at addressing unfairness often rely on sensitive user information, which can lead to unintended or intentional exploitation of data, posing privacy risks. The ML community needs more dialogue and research to grasp these privacy challenges associated with algorithmic fairness, as emphasized in studies such as [155] and [156].

1) Aligned

In this subsection, methods integrating fairness into ML models have shown positive impacts on user privacy. Dwork et al. [4] establish that individual fairness generalizes Differential Privacy (DP) under specific distance metrics, crucially enabling algorithms designed for individual fairness to enhance privacy. They define a mapping function $\mathcal{M} : X \to Y$ satisfying ϵ-DP, where distance metrics $d_{\mathcal{X}}(u, v)$ and $d_{\mathcal{Y}}(\mathcal{M}(u), \mathcal{M}(v))$ for any users u and v are defined in terms of set difference and output probability ratios. Aalmoes et al. [157] address attribute inference attacks in ML models, where adversaries attempt to infer sensitive attributes like race and gender from model outputs. They find that fairness constraints such as equalized odds effectively mitigate these attacks while preserving model utility, demonstrating alignment between fairness goals and safeguarding sensitive attribute privacy.

2) Contrasting

Chang et al. [158] define privacy risk as the susceptibility of a trained model to membership inference attacks, where adversaries use model predictions to distinguish training set members from non-members. Their study investigates the impact of fairness-aware learning on privacy risks across different demographic groups in training data, revealing a trade-off: while fairness constraints reduce bias, they increase vulnerability to membership inference attacks, particularly for underprivileged subgroups. Zhang et al. [159] explore the interaction between privacy and fairness in node classification using Graph Neural Networks (GNNs). They find that promoting individual fairness among nodes, ensuring equitable treatment regardless of background, adversely affects edge privacy measured by link prediction attacks. This empirical evidence underscores the complex relationship where efforts to enhance fairness can inadvertently compromise privacy in networked data settings.

D. Concurrent Implementation of Privacy and Fairness

This section will examine algorithms that have been proposed in existing literature with the aim of achieving both privacy and fairness objectives while minimizing the overall

loss of utility. Unfortunately, there is a limited number of research works that have theoretically investigated the interaction between these two objectives, such as those presented in references [160] and [15]. Cummings et al. [160] prove that it is not possible to simultaneously achieve DP and perfect fairness in terms of equalized odds while maintaining higher accuracy than a constant classifier. On the contrary, the authors in [15] demonstrate that this assertion is not applicable in selection problems. In non-selection problems, the goal is to minimize the expected loss across the entire population while ensuring fairness constraints. For instance, in a hiring scenario, all applicants who meet the classifier's criteria should be accepted. However, in selection problems, only a limited number of candidates can be chosen. The authors in [15] suggest that the exponential mechanism developed for DP can be an effective tool for improving fairness under specific circumstances.

Numerous studies have investigated achieving both privacy and fairness in machine learning. Liu et al. [145] proposed FairDP to address disparate impact caused by differential privacy. Chen et al. [5] examined fair classification with limited clean attributes. Xu et al. [161] combined privacy and fairness in logistic regression. Hajian et al. [162] addressed discrimination and privacy in data mining using k-anonymity. Jin et al. [163] explored privacy and fairness in inference-as-a-service. In federated learning, Padala et al. [164] introduced a framework incorporating local DP and statistical parity, while Zhang et al. [165] proposed FairFL to address information and coordination constraints.

E. Applications

Healthcare. Patient electronic health records (EHRs), containing sensitive information like gender, race, and age, require stringent legal protection. Yet, their value for research and policy-making necessitates careful balancing of privacy and fairness considerations. Synthetic datasets that mimic EHRs offer a potential solution for enhancing privacy while training equitable models [166]. However, determining the optimal degree of resemblance (privacy) and incorporating fairness notions in healthcare remains an active research area. For instance, in [167], multiple fairness notions have been suggested for synthetic healthcare data, highlighting the need for further research into the interplay between privacy and fairness in this domain.

Natural Language Processing (NLP). Modern NLP models, relying on encoded text representations, often capture sensitive attributes like race and gender, raising privacy concerns and potentially leading to unfairness. To address this, Maheshwari et al. [152] proposed a framework that perturbs text encoder outputs to achieve DP while employing an adversarial branch to promote fairness, demonstrating the synergy between these objectives. Similarly, Lyu et al. [151] introduced a noise layer for feature extraction, achieving DP before classification, further highlighting the alignment of privacy and fairness.

Computer Vision. Understanding privacy in computer vision involves various facets, with current approaches focusing on attribute privacy, where models might memorize training data or utilize features as proxies to deduce private attributes. [cite: 508] This resembles fairness concerns, as models can inadvertently transfer private information into features. Paul et al. [168] examined this privacy-fairness trade-off using an adversarial technique with penalty terms for both objectives, revealing a conflict between privacy and fairness. Conversely, Tian et al. [169] explored facial attribute classification, leveraging GANs for privacy through synthetic image generation and contrastive learning for fairness, suggesting potential for simultaneous achievement of both.

Finance. Financial decision-making is a crucial and influential domain of fairness that holds significant implications for individuals in society, while also being highly sensitive and necessitating strict privacy measures to safeguard user data [170]. Several applications that involve balancing these dual objectives include (I) Credit scoring, (II) Loan approval, (III) Insurance pricing, (IV) Target Marketing, (5) Financial advisory services, (6) Customer segmentation, and (7) Employment and promotion decisions. Legal considerations and their interplay with fairness in finance are examined and discussed in [171]. The interaction between privacy and fairness in the aforementioned applications is explored in selection problems [15] and ranking [172].

F. Synergies between privacy and fairness

1) Privacy-enhancing Fairness

Fairness-aware algorithms can indirectly enhance privacy by mitigating disparate impacts on individuals based on sensitive attributes like race, gender, or religion. For example, biased facial recognition systems can lead to misidentification and privacy violations of certain racial groups. *Reweighing* algorithms, which adjust data point weights to counter biased labels, can help mitigate these biases. Specific reweighing techniques include *demographic parity*, ensuring equal proportions of positive outcomes across groups; *equalized odds*, ensuring consistent true and false positive rates; and *calibrated equalized odds*, combining calibration and equalized odds for fair and accurate predictions. *Adversarial debiasing* is another fairness-aware technique that learns fair representations by minimizing an adversary's ability to predict sensitive attributes. This involves training models to learn group-indistinguishable representations, preventing discriminatory decision-making. Specific algorithms include *learning fair representations (LFR)*, which learns representations independent of sensitive attributes while preserving other feature information, and *variational fair autoencoder (VFAE)*, which uses a variational autoencoder to learn latent representations independent of sensitive attributes.

2) Fairness-enhancing Privacy

Privacy-preserving techniques, by limiting sensitive information, can indirectly promote fairness and prevent discrimination. DP adds noise to datasets, preventing individual identification while enabling accurate aggregate statistics. This can be achieved using mechanisms like the Laplace mechanism, which adds calibrated noise to query outputs. HE allows computations on encrypted data without decryption, preventing

Table IV: Empirical Results of Privacy and Fairness Techniques

Work	Dataset	Techniques	Measure	Results
[14]	Adult dataset	Differential Privacy	Accuracy disparity between groups	Disparity increased from 2.8% to 12.3%.
[152]	IMDB dataset	DP and Adversarial Learning	Equalized fairness, accuracy	Fairness improved from 0.62 to 0.91; accuracy $\sim$ 89%.
[139]	Synthetic dataset	Fair classification with semi-private sensitive attributes	Accuracy, fairness	Accuracy: 70% $\rightarrow$ 85%; Fairness: 0.2 $\rightarrow$ 0.8.
[153]	UCI datasets (COMPAS, Communities and Crime)	DP	Accuracy disparity between groups	Disparity increased (e.g., 10% $\rightarrow$ 15% on COMPAS).

discriminatory pattern learning during model training. Specific schemes include *BFV*, efficient for integer arithmetic; *BGV*, suitable for both integer and floating-point arithmetic; and *CKKS*, designed for approximate arithmetic in ML.

3) Effectiveness of Applying Privacy and Fairness on Empirical Data

The empirical results presented in Table IV demonstrate the effectiveness of applying privacy and fairness techniques in mitigating bias and promoting equitable outcomes in ML. While DP alone can sometimes exacerbate disparities, combining it with fairness-aware algorithms like adversarial learning can lead to significant fairness improvements with minimal impact on accuracy. This highlights the importance of integrating both privacy and fairness considerations in developing ML models to ensure responsible and equitable AI systems.

4) Available Online Tools

The increasing importance of algorithmic privacy and fairness has spurred the development of numerous online tools to facilitate the integration of ethical considerations into machine learning models. Notable examples include *Fairlearn* [173] , developed by Microsoft Research for assessing and improving fairness; *AI Fairness 360* (AIF360) [174], an open-source toolkit by IBM providing metrics and mitigation algorithms for addressing bias; and *Aequitas* [175], which offers auditing capabilities for identifying and eliminating bias. Furthermore, tools like Google's *What-If Tool* [176] and LinkedIn Fairness Toolkit (LiFT) [177] contribute to enhancing algorithmic fairness. These tools empower practitioners to build and deploy responsible AI systems that prioritize both privacy and fairness.

V. Vision and Challenges

A. Privacy and Fairness in the Current Era of Large Language Models

In the evolving landscape of large language models (LLMs), the dual imperatives of privacy and fairness are increasingly paramount, particularly as these models become more interactive and pervasive. APIs facilitating access to LLMs, exemplified by platforms like OpenAI [178], [179] and others [180], incorporate rigorous checks to ensure both the quality of model outputs and adherence to fairness principles. Logic-aware models [181] represent a significant stride in bias mitigation by employing textual entailment techniques, thereby promoting fairness without necessitating additional data sources. Moreover, the integration of human feedback in fine-tuning LLMs [182] not only refines model behavior to align with user preferences but also underscores efforts to maintain privacy during training and deployment. These multifaceted approaches underscore the ongoing quest to harmonize technological advancement with ethical considerations, crucial for fostering public trust and acceptance of AI technologies.

B. Fairness Through Privacy

The majority of previous approaches aimed at mitigating bias require access to sensitive attributes. However, obtaining a significant amount of data with sensitive attributes is often impractical due to people's growing privacy concerns and legal compliance. Consequently, a crucial area of research inquiry that merits attention is how to ensure fair predictions while preserving privacy. This is a persistent challenge faced by technology companies that seek to balance the goal of ensuring fair ML processing of user data, including sensitive attributes such as Race and Gender, while simultaneously protecting user privacy and restricting the use of sensitive user data.

C. Fair Privacy Protection

The concept of fair privacy protection, as discussed in [183], questions whether privacy mechanisms provide equal protection across different user groups. While the goal is to ensure fairness by offering uniform privacy levels to all, certain groups may receive disproportionate attention in practice. For instance, the US Census Bureau's use of the Laplace mechanism in DP illustrates this disparity. Low-population areas like villages may receive higher noise additions per individual compared to densely populated cities under the same ϵ-DP constraint [184]. This discrepancy arises because noise variance per individual scales inversely with population size, potentially compromising the fairness of privacy guarantees across diverse demographic settings.

This disparity highlights the need for more nuanced and equitable privacy-preserving mechanisms that account for the diverse characteristics of different population groups. Future research should focus on developing adaptive privacy mechanisms that adjust the level of protection based on factors like population density, demographic composition, and sensitivity of the data. Additionally, group-aware privacy metrics should be developed to explicitly measure and monitor privacy protection at the group level rather than solely relying on individual-

level metrics. Furthermore, fairness-aware data release mechanisms should be designed to ensure that the dissemination of information does not disproportionately benefit or harm any particular group while still preserving privacy.

D. Incorporating Privacy and Fairness based on Cryptographic Approaches

No existing approach addresses *both* privacy and fairness in the cryptographic setting. Such an approach presents great promise, because it may be able to provide privacy and fairness under more relaxed system architecture assumptions. For instance, in the differential privacy case, one assumes the presence of a trusted curator, or that an extensive distributed infrastructure for federated learning exists. With cryptography, no trusted party is required to perform the computation.

The main challenge becomes how can one express fairness constraints so that they become implementable using the rather restrictive set of operations provided by various searchable encryption approaches. Can one achieve fairness directly under the encrypted ciphertext using primitives like PHE? Or are there more expensive primitives required, like FHE? And even with FHE, only polynomial evaluation is supported in the best case, whereas other operations (e.g., logarithm, sigmoid) must be simulated using polynomial approximations. Achieving fairness by using such primitives is an important and challenging research problem.

VI. Conclusion

In conclusion, this comprehensive survey offers a thorough investigation of the fundamental concepts in privacy and fairness by examining nearly 200 works in the field. Our aim is to guide researchers in both academia and industry towards the simultaneous realization of privacy and fairness for individuals and groups in society across all four primary facets of ML, including supervised, unsupervised, semi-supervised, and reinforcement learning. By establishing a solid understanding of privacy and fairness within various ML techniques, we present an exhaustive analysis of how objectives impact one another and identify open questions for the first time. This work emphasizes the focus areas necessary to address the dual objectives in ML, ultimately promoting more responsible and trustworthy decision-making.

References

[1] H. Surden, "Machine learning and law," *Wash. L. Rev.*, vol. 89, p. 87, 2014.

[2] A. Chalfin, O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan, "Productivity and selection of human capital with machine learning," *American Economic Review*, vol. 106, no. 5, pp. 124–127, 2016.

[3] F. Hussain, S. A. Hassan, R. Hussain, and E. Hossain, "Machine learning for resource management in cellular and iot networks: Potentials, current solutions, and open challenges," *IEEE communications surveys & tutorials*, vol. 22, no. 2, pp. 1251–1275, 2020.

[4] 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*, 2012, pp. 214–226.

[5] C. Chen, Y. Liang, X. Xu, S. Xie, Y. Hong, and K. Shu, "When fairness meets privacy: Fair classification with semi-private sensitive attributes," in *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

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

[7] D. J. Solove, "Understanding privacy," 2008.

[8] F. Times, "Facebook privacy breach," *Financial Times*, pp. 11–12, 2020. [Online]. Available: https://www.ft.com/content/87184c402cfe-11e8-9b4b-bc4b9f08f381

[9] F. T. Council. (2023) How privacy got on the calendar. Accessed: 2023-05-01. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2023/01/24/how-privacy-got-on-the-calendar/?sh=517cc83959f7

[10] T. H. Anderson, *The pursuit of fairness: A history of affirmative action*. Oxford University Press, 2004.

[11] J. Angwin, J. Larson, L. Kirchner, and S. Mattu, "Machine bias," May 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[12] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 469–481.

[13] N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1083–1094, 2022.

[14] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," *Advances in Neural Information Processing Systems*, vol. 32, pp. 15 479–15 488, 2019.

[15] M. M. Khalili, X. Zhang, M. Abroshan, and S. Sojoudi, "Improving fairness and privacy in selection problems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 8092–8100.

[16] M. Pannekoek and G. Spigler, "Investigating trade-offs in utility, fairness and differential privacy in neural networks," *arXiv preprint arXiv:2102.05975*, 2021.

[17] E. U. Soykan, L. Karaçay, F. Karakoç, and E. Tomur, "A survey and guideline on privacy enhancing technologies for collaborative machine learning," *IEEE Access*, vol. 10, pp. 97 495–97 519, 2022.

[18] A. Blanco-Justicia, D. Sanchez, J. Domingo-Ferrer, and K. Muralidhar, "A critical review on the use (and misuse) of differential privacy in machine learning," *arXiv preprint arXiv:2206.04621*, 2022.

[19] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1452, 2022.

[20] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[21] M. Choudhary, C. Laclau, and C. Largeron, "A survey on fairness for machine learning on graphs," *arXiv preprint arXiv:2205.05396*, 2022.

[22] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–36, 2021.

[23] E. De Cristofaro, "A critical overview of privacy in machine learning," *IEEE Security & Privacy*, vol. 19, no. 4, pp. 19–27, 2021.

[24] R. Xu, N. Baracaldo, and J. Joshi, "Privacy-preserving machine learning: Methods, challenges and directions," *arXiv preprint arXiv:2108.04417*, 2021.

[25] M. Wan, D. Zha, N. Liu, and N. Zou, "Modeling techniques for machine learning fairness: A survey," *arXiv preprint arXiv:2111.03015*, 2021.

[26] A. Chhabra, K. Masalkovaitė, and P. Mohapatra, "An overview of fairness in clustering," *IEEE Access*, 2021.

[27] H. C. Tanuwidjaja, R. Choi, S. Baek, and K. Kim, "Privacy-preserving deep learning on machine learning as a service—a comprehensive survey," *IEEE Access*, vol. 8, pp. 167 425–167 447, 2020.

[28] A. Nayyar, L. Gadhavi, and N. Zaman, "Machine learning in healthcare: review, opportunities and challenges," *Machine Learning and the Internet of Medical Things in Healthcare*, pp. 23–45, 2021.

[29] A. Tizghadam, H. Khazaei, M. H. Moghaddam, and Y. Hassan, "Machine learning in transportation," 2019.

[30] M. F. Dixon, I. Halperin, and P. Bilokon, *Machine learning in Finance*. Springer, 2020, vol. 1170.

[31] K. Y. Ngiam and W. Khor, "Big data and machine learning algorithms for health-care delivery," *The Lancet Oncology*, vol. 20, no. 5, pp. e262–e273, 2019.

[32] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[33] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryption schemes: Theory and implementation," *ACM Computing Surveys (Csur)*, vol. 51, no. 4, pp. 1–35, 2018.

[34] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[35] N. Holohan, S. Antonatos, S. Braghin, and P. Mac Aonghusa, "The bounded laplace mechanism in differential privacy," *arXiv preprint arXiv:1808.10410*, 2018.

[36] F. Koufogiannis, S. Han, and G. J. Pappas, "Optimality of the laplace mechanism in differential privacy," *arXiv preprint arXiv:1504.00065*, 2015.

[37] N. Fernandes, A. McIver, and C. Morgan, "The laplace mechanism has optimal utility for differential privacy over continuous queries," in *2021 36th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*. IEEE, 2021, pp. 1–12.

[38] V. Rastogi and S. Nath, "Differentially private aggregation of distributed time-series with transformation and encryption," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 2010, pp. 735–746.

[39] S. Shaham, G. Ghinita, and C. Shahabi, "Differentially-private publication of origin-destination matrices with intermediate stops," *arXiv preprint arXiv:2202.12342*, 2022.

[40] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, "Differentially private histogram publication," *The VLDB journal*, vol. 22, pp. 797–822, 2013.

[41] S. Shaham, G. Ghinita, R. Ahuja, J. Krumm, and C. Shahabi, "Htf: Homogeneous tree framework for differentially-private release of large geospatial datasets with self-tuning structure height," *ACM Transactions on Spatial Algorithms and Systems*, 2022.

[42] ——, "Htf: Homogeneous tree framework for differentially-private release of location data," in *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, 2021, pp. 184–194.

[43] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," *IEEE Transactions on knowledge and data engineering*, vol. 23, no. 8, pp. 1200–1214, 2010.

[44] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu, "Differentially private spatial decompositions," in *2012 IEEE 28th International Conference on Data Engineering*. IEEE, 2012, pp. 20–31.

[45] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 2007, pp. 94–103.

[46] J. Dong, D. Durfee, and R. Rogers, "Optimal differential privacy composition for exponential mechanisms," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2597–2606.

[47] D. Feldman, A. Fiat, H. Kaplan, and K. Nissim, "Private coresets," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009, pp. 361–370.

[48] B. I. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, "Learning in a large function space: Privacy-preserving mechanisms for svm learning," *arXiv preprint arXiv:0911.5708*, 2009.

[49] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering frequent patterns in sensitive data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 503–512.

[50] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[51] V. M. Suriyakumar, N. Papernot, A. Goldenberg, and M. Ghassemi, "Challenges of differentially private prediction in healthcare settings," in *Proceedings of the IJCAI 2021 Workshop on AI for Social Good*, 2021.

[52] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[53] H. B. McMahan, G. Andrew, U. Erlingsson, S. Chien, I. Mironov, N. Papernot, and P. Kairouz, "A general approach to adding differential privacy to iterative training procedures," *arXiv preprint arXiv:1812.06210*, 2018.

[54] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao *et al.*, "Opacus: User-friendly differential privacy library in pytorch," *arXiv preprint arXiv:2109.12298*, 2021.

[55] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, pp. 45–66, 2004.

[56] J. Park, Y. Choi, J. Byun, J. Lee, and S. Park, "Efficient differentially private kernel support vector classifier for multi-class classification," *Information Sciences*, vol. 619, pp. 889–907, 2023.

[57] X. Yi, R. Paulet, E. Bertino, X. Yi, R. Paulet, and E. Bertino, *Homomorphic encryption*. Springer, 2014.

[58] D. Micciancio and O. Regev, "Lattice-based cryptography," *Post-quantum cryptography*, pp. 147–191, 2009.

[59] G. Hanrot, X. Pujol, and D. Stehlé, "Algorithms for the shortest and closest lattice vector problems." *IWCC*, vol. 6639, pp. 159–190, 2011.

[60] H. Yousuf, M. Lahzi, S. A. Salloum, and K. Shaalan, "Systematic review on fully homomorphic encryption scheme and its application," *Recent Advances in Intelligent Systems and Smart Applications*, pp. 537–551, 2020.

[61] M. Ogburn, C. Turner, and P. Dahal, "Homomorphic encryption," *Procedia Computer Science*, vol. 20, pp. 502–509, 2013.

[62] K. G. Kogos, K. S. Filippova, and A. V. Epishkina, "Fully homomorphic encryption schemes: The state of the art," in *2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE, 2017, pp. 463–466.

[63] C. Moore, M. O'Neill, E. O'Sullivan, Y. Doröz, and B. Sunar, "Practical homomorphic encryption: A survey," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2014, pp. 2792–2795.

[64] J. Sen, "Homomorphic encryption-theory and application," *Theory and practice of cryptography and network security protocols and technologies*, vol. 31, 2013.

[65] A. Yu, A. V. Sathanur, and V. Jandhyala, "A partial homomorphic encryption scheme for secure design automation on public clouds," in *2012 IEEE 21st Conference on Electrical Performance of Electronic Packaging and Systems*. IEEE, 2012, pp. 177–180.

[66] D. P. Hellwig and A. Huchzermeier, "Distributed ledger technology and fully homomorphic encryption: Next-generation information-sharing for supply chain efficiency," in *Innovative Technology at the Interface of Finance and Operations: Volume II*. Springer, 2022, pp. 31–49.

[67] J. Fan and F. Vercauteren, "Somewhat practical fully homomorphic encryption," *Cryptology ePrint Archive*, 2012.

[68] D. Boneh, C. Gentry, S. Halevi, F. Wang, and D. J. Wu, "Private database queries using somewhat homomorphic encryption," in *Applied Cryptography and Network Security: 11th International Conference, ACNS 2013, Banff, AB, Canada, June 25-28, 2013. Proceedings 11*. Springer, 2013, pp. 102–118.

[69] C. Aguilar-Melchor, S. Fau, C. Fontaine, G. Gogniat, and R. Sirdey, "Recent advances in homomorphic encryption: A possible future for signal processing in the encrypted domain," *IEEE Signal Processing Magazine*, vol. 30, no. 2, pp. 108–117, 2013.

[70] R. Hamza, A. Hassan, A. Ali, M. B. Bashir, S. M. Alqhtani, T. M. Tawfeeg, and A. Yousif, "Towards secure big data analysis via fully homomorphic encryption algorithms," *Entropy*, vol. 24, no. 4, p. 519, 2022.

[71] V. Migliore, G. Bonnoron, and C. Fontaine, "Practical parameters for somewhat homomorphic encryption schemes on binary circuits," *IEEE Transactions on Computers*, vol. 67, no. 11, pp. 1550–1560, 2018.

[72] C. D. Wenger, D. H. Phanstiel, M. V. Lee, D. J. Bailey, and J. J. Coon, "Compass: A suite of pre-and post-search proteomics software tools for omssa," *Proteomics*, vol. 11, no. 6, pp. 1064–1074, 2011.

[73] A. Campolo, M. R. Sanfilippo, M. Whittaker, and K. Crawford, "Ai now 2017 report," 2017.

[74] M. J. Zimmer, "Emerging uniform structure of disparate treatment discrimination litigation," *Ga. L. Rev.*, vol. 30, p. 563, 1995.

[75] G. Rutherglen, "Disparate impact under title vii: an objective theory of discrimination," *Va. L. Rev.*, vol. 73, p. 1297, 1987.

[76] A. D. Selbst, "Disparate impact in big data policing," *Ga. L. Rev.*, vol. 52, p. 109, 2017.

[77] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data mining and knowledge discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[78] A. W. Flores, K. Bechtel, and C. T. Lowenkamp, "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks," *Fed. Probation*, vol. 80, p. 38, 2016.

[79] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *Advances in neural information processing systems*, vol. 30, 2017.

[80] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in neural information processing systems*, vol. 30, 2017.

[81] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[82] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi, "Fairness without demographics through adversarially reweighted learning," *Advances in neural information processing systems*, vol. 33, pp. 728–740, 2020.

[83] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Fairbatch: Batch selection for model fairness," *arXiv preprint arXiv:2012.01696*, 2020.

[84] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[85] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[86] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015.

[87] B. T. Luong, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 502–510.

[88] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2012, pp. 35–50.

[89] Y. Bechavod and K. Ligett, "Penalizing unfairness in binary classification," *arXiv preprint arXiv:1707.00044*, 2017.

[90] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," *arXiv preprint arXiv:1706.02409*, 2017.

[91] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *2010 IEEE international conference on data mining*. IEEE, 2010, pp. 869–874.

[92] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 962–970.

[93] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[94] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning non-discriminatory predictors," in *Conference on Learning Theory*. PMLR, 2017, pp. 1920–1953.

[95] N. Quadrianto and V. Sharmanska, "Recycling privileged learning and distribution matching for fairness," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[96] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[97] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris, "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 853–862.

[98] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[99] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers," in *Icml*, vol. 1. Citeseer, 2001, pp. 609–616.

[100] S. Shaham, G. Ghinita, and C. Shahabi, "Models and mechanisms for spatial data fairness," *Proceedings of the VLDB Endowment*, vol. 16, no. 2, pp. 167–179, 2022.

[101] F. Petersen, D. Mukherjee, Y. Sun, and M. Yurochkin, "Post-processing for individual fairness," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 944–25 955, 2021.

[102] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 247–254.

[103] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 924–929.

[104] A. K. Menon and R. C. Williamson, "The cost of fairness in classification," *arXiv preprint arXiv:1705.09055*, 2017.

[105] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[106] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.

[107] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, "Bias mitigation post-processing for individual and group fairness," in *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2019, pp. 2847–2851.

[108] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Fair clustering through fairlets," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[109] S. Bera, D. Chakrabarty, N. Flores, and M. Negahbani, "Fair algorithms for clustering," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[110] S. Ahmadian, A. Epasto, R. Kumar, and M. Mahdian, "Clustering without over-representation," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 267–275.

[111] A. Chhabra, V. Vashishth, and P. Mohapatra, "Fair algorithms for hierarchical agglomerative clustering," *arXiv preprint arXiv:2005.03197*, 2020.

[112] M. Ghadiri, S. Samadi, and S. Vempala, "Socially fair k-means clustering," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 438–448.

[113] M. Kleindessner, P. Awasthi, and J. Morgenstern, "A notion of individual fairness for clustering," *arXiv preprint arXiv:2006.04960*, 2020.

[114] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Fair clustering through fairlets," *Advances in neural information processing systems*, vol. 30, 2017.

[115] S. Ahmadian, A. Epasto, M. Knittel, R. Kumar, M. Mahdian, B. Moseley, P. Pham, S. Vassilvitskii, and Y. Wang, "Fair hierarchical clustering," *arXiv preprint arXiv:2006.10221*, 2020.

[116] B. Rastegarpanah, K. P. Gummadi, and M. Crovella, "Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems," in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 231–239.

[117] A. Chhabra, A. Singla, and P. Mohapatra, "Fair clustering using antidote data," in *Algorithmic Fairness through the Lens of Causality and Robustness workshop*. PMLR, 2022, pp. 19–39.

[118] M. Kleindessner, S. Samadi, P. Awasthi, and J. Morgenstern, "Guarantees for spectral clustering with fairness constraints," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3458–3467.

[119] P. Li, H. Zhao, and H. Liu, "Deep fair clustering for visual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9070–9079.

[120] H. Zhang and I. Davidson, "Deep fair discriminative clustering," *arXiv preprint arXiv:2105.14146*, 2021.

[121] J. Chai and X. Wang, "Fair clustering via equalized confidence."

[122] S. Liu and L. N. Vicente, "A stochastic alternating balance k-means algorithm for fair clustering," in *Learning and Intelligent Optimization: 16th International Conference, LION 16, Milos Island, Greece, June 5–10, 2022, Revised Selected Papers*. Springer, 2023, pp. 77–92.

[123] I. M. Ziko, J. Yuan, E. Granger, and I. B. Ayed, "Variational fair clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 11 202–11 209.

[124] X. Chen, B. Fain, L. Lyu, and K. Munagala, "Proportionally fair clustering," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1032–1041.

[125] S. Simoes, P. Deepak, and M. MacCarthaigh, "Exploring rawlsian fairness for k-means clustering," in *Responsible Data Science: Select Proceedings of ICDSE 2021*. Springer, 2022, pp. 47–59.

[126] J. Altham, "Rawls's difference principle," *Philosophy*, vol. 48, no. 183, pp. 75–78, 1973.

[127] E. Harb and H. S. Lam, "Kfc: A scalable approximation algorithm for k- center fair clustering," *Advances in neural information processing systems*, vol. 33, pp. 14 509–14 519, 2020.

[128] S. Esmaeili, B. Brubach, L. Tsepenekas, and J. Dickerson, "Probabilistic fair clustering," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 743–12 755, 2020.

[129] S. Esmaeili, B. Brubach, A. Srinivasan, and J. Dickerson, "Fair clustering under a bounded cost," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 345–14 357, 2021.

[130] Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma, "A survey on the fairness of recommender systems," *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1–43, 2023.

[131] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[132] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth, "Fairness in learning: Classic and contextual bandits," *Advances in neural information processing systems*, vol. 29, 2016.

[133] Y. Liu, G. Radanovic, C. Dimitrakakis, D. Mandal, and D. C. Parkes, "Calibrated fairness in bandits," *arXiv preprint arXiv:1707.01875*, 2017.

[134] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1799–1813, 2019.

[135] H. Claure, Y. Chen, J. Modi, M. Jung, and S. Nikolaidis, "Multi-armed bandits with fairness constraints for distributing resources to human teammates," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 299–308.

[136] V. Patil, G. Ghalme, V. Nair, and Y. Narahari, "Achieving fairness in the stochastic multi-armed bandit problem," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 7885–7915, 2021.

[137] W. Liu, F. Liu, R. Tang, B. Liao, G. Chen, and P. A. Heng, "Balancing between accuracy and fairness for interactive recommendation with reinforcement learning," in *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*. Springer, 2020, pp. 155–167.

[138] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, "Fairness in reinforcement learning," in *International conference on machine learning*. PMLR, 2017, pp. 1617–1626.

[139] Y. Chen, A. Cuellar, H. Luo, J. Modi, H. Nemlekar, and S. Nikolaidis, "Fair contextual multi-armed bandits: Theory and experiments," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 181–190.

[140] W. Huang, K. Labille, X. Wu, D. Lee, and N. Heffernan, "Achieving user-side fairness in contextual bandits," *Human-Centric Intelligent Systems*, pp. 1–14, 2022.

[141] M. Wen, O. Bastani, and U. Topcu, "Algorithms for fairness in sequential decision making," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1144–1152.

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

[143] Y. Ge, X. Zhao, L. Yu, S. Paul, D. Hu, C.-C. Hsieh, and Y. Zhang, "Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning," in *Proceedings of the fifteenth ACM international conference on web search and data mining*, 2022, pp. 316–324.

[144] D. Su, J. Cao, N. Li, E. Bertino, and H. Jin, "Differentially private k-means clustering," in *Proceedings of the sixth ACM conference on data and application security and privacy*, 2016, pp. 26–37.

[145] W. Liu, X. Wang, X. Lu, J. Cheng, B. Jin, X. Wang, and H. Zha, "Fair differential privacy can mitigate the disparate impact on model accuracy," 2020.

[146] A. Lowy, D. Gupta, and M. Razaviyayn, "Stochastic differentially private and fair learning," *arXiv preprint arXiv:2210.08781*, 2022.

[147] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.

[148] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and S. Avestimehr, "Fairfed: Enabling group fairness in federated learning," *arXiv preprint arXiv:2110.00857*, 2021.

[149] S. Pentyala, D. Melanson, M. De Cock, and G. Farnadi, "Privfair: a library for privacy-preserving fairness auditing," *arXiv preprint arXiv:2202.04058*, 2022.

[150] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni, "On the fairness of privacy-preserving representations in medical applications," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*. Springer, 2020, pp. 140–149.

[151] L. Lyu, X. He, and Y. Li, "Differentially private representation for nlp: Formal guarantee and an empirical study on privacy and fairness," *arXiv preprint arXiv:2010.01285*, 2020.

[152] G. Maheshwari, P. Denis, M. Keller, and A. Bellet, "Fair nlp models with differentially private text encoders," *arXiv preprint arXiv:2205.06135*, 2022.

[153] A. Sanyal, Y. Hu, and F. Yang, "How unfair is private learning?" in *Uncertainty in Artificial Intelligence*. PMLR, 2022, pp. 1738–1748.

[154] M. Du, R. Jia, and D. Song, "Robust anomaly detection and backdoor attack detection via differential privacy," *arXiv preprint arXiv:1911.07116*, 2019.

[155] M. Andrus and S. Villeneuve, "Demographic-reliant algorithmic fairness: characterizing the risks of demographic data collection in the pursuit of fairness," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1709–1721.

[156] M. Strobel and R. Shokri, "Data privacy and trustworthy machine learning," *IEEE Security & Privacy*, vol. 20, no. 5, pp. 44–49, 2022.

[157] J. Aalmoes, V. Duddu, and A. Boutet, "Leveraging algorithmic fairness to mitigate blackbox attribute inference attacks," *arXiv preprint arXiv:2211.10209*, 2022.

[158] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021, pp. 292–303.

[159] H. Zhang, X. Yuan, Q. V. H. Nguyen, and S. Pan, "On the interaction between node fairness and edge privacy in graph neural networks," *arXiv preprint arXiv:2301.12951*, 2023.

[160] 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*, 2019, pp. 309–315.

[161] D. Xu, S. Yuan, and X. Wu, "Achieving differential privacy and fairness in logistic regression," in *Companion proceedings of The 2019 world wide web conference*, 2019, pp. 594–599.

[162] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti, "Discrimination-and privacy-aware patterns," *Data Mining and Knowledge Discovery*, vol. 29, no. 6, pp. 1733–1782, 2015.

[163] Y. Jin and L. Lai, "Privacy protection in learning fair representations," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2964–2968.

[164] M. Padala, S. Damle, and S. Gujar, "Federated learning meets fairness and differential privacy," in *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part VI 28*. Springer, 2021, pp. 692–699.

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

[166] K. P. Seastedt, P. Schwab, Z. O'Brien, E. Wakida, K. Herrera, P. G. F. Marcelo, L. Agha-Mir-Salim, X. B. Frigola, E. B. Ndulue, A. Marcelo *et al.*, "Global healthcare fairness: We should be sharing more, not less, data," *PLOS Digital Health*, vol. 1, no. 10, p. e0000102, 2022.

[167] K. Bhanot, M. Qi, J. S. Erickson, I. Guyon, and K. P. Bennett, "The problem of fairness in synthetic healthcare data," *Entropy*, vol. 23, no. 9, p. 1165, 2021.

[168] W. Paul, P. Mathew, F. Alajaji, and P. Burlina, "Evaluating trade-offs in computer vision between attribute privacy, fairness and utility," *arXiv preprint arXiv:2302.07917*, 2023.

[169] H. Tian, T. Zhu, and W. Zhou, "Fairness and privacy preservation for facial images: Gan-based methods," *Computers & Security*, vol. 122, p. 102902, 2022.

[170] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[171] S. Das, M. Donini, J. Gelman, K. Haas, M. Hardt, J. Katzman, K. Kenthapadi, P. Larroy, P. Yilmaz, and B. Zafar, "Fairness measures for machine learning in finance," 2021.

[172] J. A. Sun, S. Pentyala, M. D. Cock, and G. Farnadi, "Privacy-preserving fair item ranking," in *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*. Springer, 2023, pp. 188–203.

[173] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in ai," *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[174] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[175] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.

[176] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 56–65, 2019.

[177] S. Vasudevan and K. Kenthapadi, "Lift: A scalable framework for measuring fairness in ml applications," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2773–2780.

[178] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[179] OpenAI, "Gpt-4 technical report," 2023.

[180] S. Soltan, S. Ananthakrishnan, J. G. M. FitzGerald, R. Gupta, W. Hamza, H. Khan, C. Peris, S. Rawls, A. Rosenbaum, A. Rumshisky, C. S. Prakash, M. Sridhar, F. Triefenbach, A. Verma, G. Tur, and P. Natarajan, "Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model," *arXiv*, 2022. [Online]. Available: https://www.amazon.science/publications/alexatm-20b-few-shot-learning-using-a-large-scale-multilingual-seq2seq-model

[181] H. Luo and J. Glass, "Logic against bias: Textual entailment mitigates stereotypical sentence reasoning," 2023.

[182] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022.

[183] M. D. Ekstrand, R. Joshaghani, and H. Mehrpouyan, "Privacy for all: Ensuring fair and equitable privacy protections," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 35–47.

[184] W. P. O'Hare, *Differential undercounts in the US census: who is missed?* Springer Nature, 2019.