Computational Intelligence

WILEY

Computational Intelligence

SPECIAL ISSUE ARTICLE OPEN ACCESS

Taming the Triangle: On the Interplays Between Fairness, Interpretability, and Privacy in Machine Learning

Julien Ferry[1] | Ulrich Aïvodji[2] | Sébastien Gambs[3] | Marie-José Huguet[4] | Mohamed Siala[4]

[1]CIRRELT & SCALE-AI Chair in Data-Driven Supply Chains, Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal, Canada | [2]École de Technologie Supérieure, Montréal, Canada | [3]Université du Québec à Montréal, Montréal, Canada | [4]LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France

Correspondence: Julien Ferry (julien.ferry@polymtl.ca)

Received: 31 October 2024 | **Revised:** 23 June 2025 | **Accepted:** 15 July 2025

Funding: This work was supported by Polytechnique Montréal, the LabEx CIMI (Grant/Award Number: ANR-11-LABX-0040) and the Canada Research Chairs Program (Privacy-preserving and ethical analysis of Big Data Canada Research Chair).

Keywords: explainability | fairness | interpretability | machine learning | privacy

ABSTRACT

Machine learning techniques are increasingly used for high-stakes decision-making, such as college admissions, loan attribution, or recidivism prediction. Thus, it is crucial to ensure that the models learnt can be audited or understood by human users, do not create or reproduce discrimination or bias and do not leak sensitive information regarding their training data. Indeed, interpretability, fairness, and privacy are key requirements for the development of responsible machine learning, and all three have been studied extensively during the last decade. However, they were mainly considered in isolation, while in practice they interplay with each other, either positively or negatively. In this survey paper, we review the literature on the interactions between these three desiderata. More precisely, for each pairwise interaction, we summarize the identified synergies and tensions. These findings highlight several fundamental theoretical and empirical conflicts, while also demonstrating that jointly considering these different requirements is challenging when one aims at preserving a high level of utility. To solve this issue, we also discuss possible conciliation mechanisms, showing that a careful design can enable to successfully handle these different concerns in practice.

1 | Introduction

Machine learning (ML) models have many useful and promising applications. For instance, they can help to analyze medical data, which is becoming increasingly complex due to the improvements in medical tools. However, their growing use for high-stakes decision-making systems—such as college admissions, recidivism prediction or credit scoring—raises significant ethical, philosophical, and societal challenges. This has led to the regulation of their use through legislations, such as the European Union General Data Protection Regulation (GDPR)[1] [1] or the AI Act[2].

In particular, three important ethical issues have emerged, each corresponding to a key concern that should be addressed to both comply with these new legal frameworks and lay the foundations toward a responsible ML. First, ML algorithms require large amounts of data, which often contains personal information. Thus, it is of paramount importance to ensure that the *privacy* of the involved individuals is not harmed while also being able to extract useful generic patterns from this data. Regulatory frameworks also directly mandate such data protection—most notably the GDPR, but also the AI Act, for instance through Article 10 on data and data governance. Second, it was shown that data-driven decision-making processes can create or

reproduce biases that systematically disadvantage specific individuals or groups [2]. Quantifying but also reducing/eliminating these biases to promote *fairness* is hence an important challenge, targeted by recent AI regulations and also closely related to pre-existing non-discrimination laws [3]. Third, while common ML models, such as deep neural networks, can reach high predictive performance, their underlying logic and representation are often too complex, preventing users from fully understanding their decisions. This raises significant concerns, regarding their auditability, certifiability, and trust, thus calling for the requirement of *interpretability* with respect to their predictions. For example, the AI Act mandates transparency for all "high-risk" AI systems—which in practice covers many applications—under Article 13. Additionally, Article 86 introduces a so-called "right to explanation" of individual decision-making.

These three topics, namely privacy, fairness, and explainability, have been extensively studied during the past decade [4–6] with an emphasis on how they each trade-off with utility. However, they are often considered in isolation, while in practice it is necessary to enforce them *simultaneously*. Characterizing their mutual interplays is hence an important research avenue, which has attracted some attention in the last years. Indeed, these concerns often conflict [7], and trade-offs between them, as well as with utility, generally have to be set. Throughout this survey paper, we conduct an in-depth review of the literature on the different compatibilities, synergies and tensions that have been identified between them. More precisely, we focus on the supervised learning setup while considering mainly classification tasks.

Despite growing regulatory pressure, the real-world deployment of techniques designed to meet the aforementioned ethical desiderata often encounters substantial practical challenges. For example, the use of Differential Privacy (DP) by the United States Census Bureau for the 2020 Census of Population and Housing offers valuable insights into the operational difficulties organizations may face when implementing DP in practice [8]. Similarly, algorithmic fairness approaches may fail to align with non-discrimination regulations—which often require case-by-case assessments—and can even come into conflict in certain contexts [9, 10]. These challenges are only exacerbated when considering jointly the different ethical desiderata. For instance, the use of DP for the release of the 2020 U.S. Census of Population and Housing provided thorough privacy protection to the released data. However, the noise added to preserve privacy was later shown to yield unfair outcomes for downstream resource allocation tasks, resulting for instance in disparate budget allocation for school districts [11, 12]. The Apple Card controversy provides another striking real-world example. More precisely in 2019, users and journalists raised concerns that the card's credit limit algorithm may assign significantly lower credit limits to women, even when they had similar or better financial profiles than men [13]. Notably, the algorithm did not use gender as an explicit input attribute—a design choice intended to protect user privacy and promote fairness. However, this omission complicated efforts to audit and correct potential biases, as gender-related disparities could still arise indirectly through correlated attributes. This case illustrates the ongoing tension between protecting the privacy of sensitive attributes and enabling effective fairness audits and enforcement [14].

1.1 | Positioning With Respect to Other Surveys

Other recent works survey the literature on the interactions between several of our three identified desiderata. Among others [7], review at a high level the main tensions that occur between the human values of privacy, transparency, and fairness when they have to be embodied in a machine learning model. We extend this work by additionally considering compatibilities and synergies. Furthermore, while they also discuss tensions within each pillar and with the context of deployment, we rather focus on the interplays between the three aspects to allow a more thorough technical discussion. Furthermore [12], investigate solely the interplays between fairness and (differential) privacy by conducting an in-depth analysis on how one influences the other. We extend this study in Section 4 [15] focuses solely on the federated learning setting and surveys approaches aimed at privacy

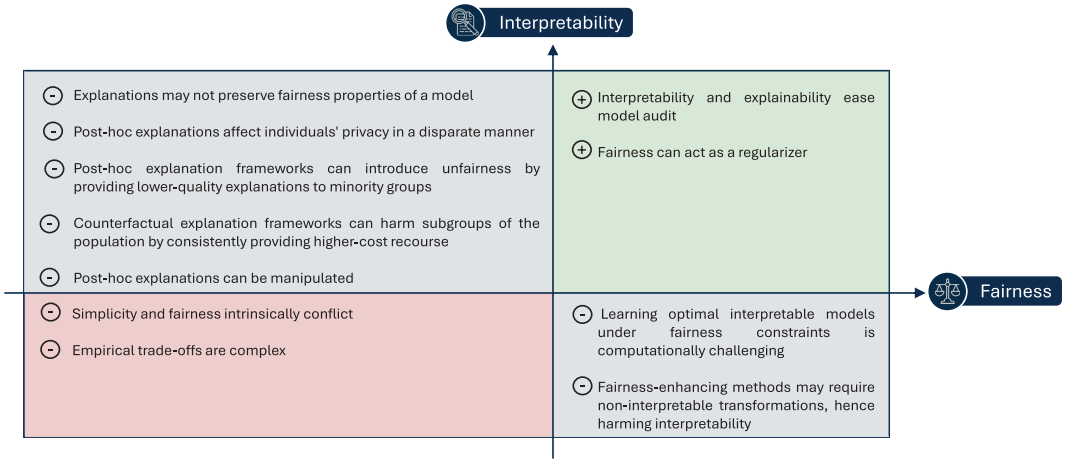

FIGURE 1 | Summary of the identified tensions (−) (Section 3.1) and compatibilities/synergies (+) (Section 3.2) between fairness and interpretability in machine learning. Fairness is represented conceptually along the x-axis: observations on the left tend to harm fairness, while those on the right tend to enhance it. Similarly, interpretability is represented on the y-axis: observations near the bottom tend to harm interpretability or explainability, while those near the top tend to enhance it.

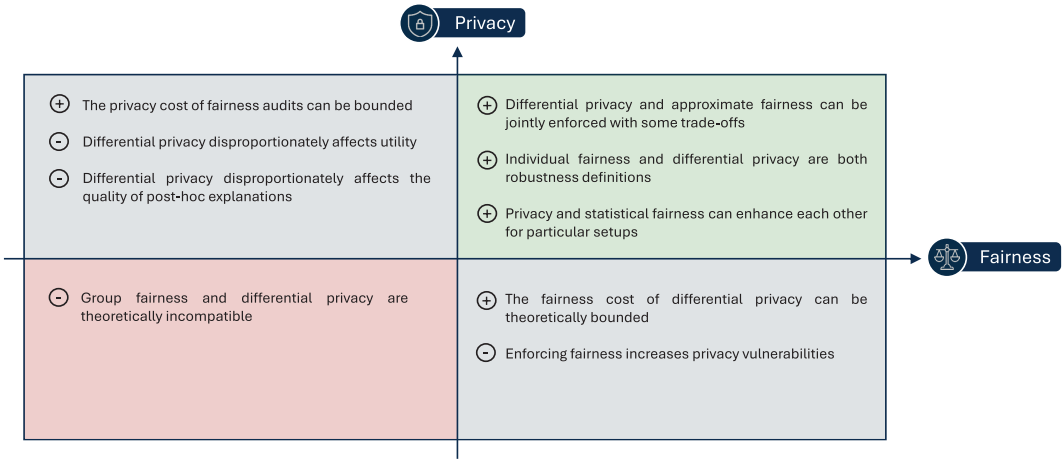

FIGURE 2 | Summary of the identified tensions (−) (Section 4.1) and compatibilities/synergies (+) (Section 4.2) between fairness and privacy in machine learning. Fairness is represented conceptually along the x-axis: observations on the left tend to harm fairness, while those on the right tend to enhance it. Similarly, privacy is represented on the y-axis: observations near the bottom tend to harm privacy, while those near the top tend to enhance it.

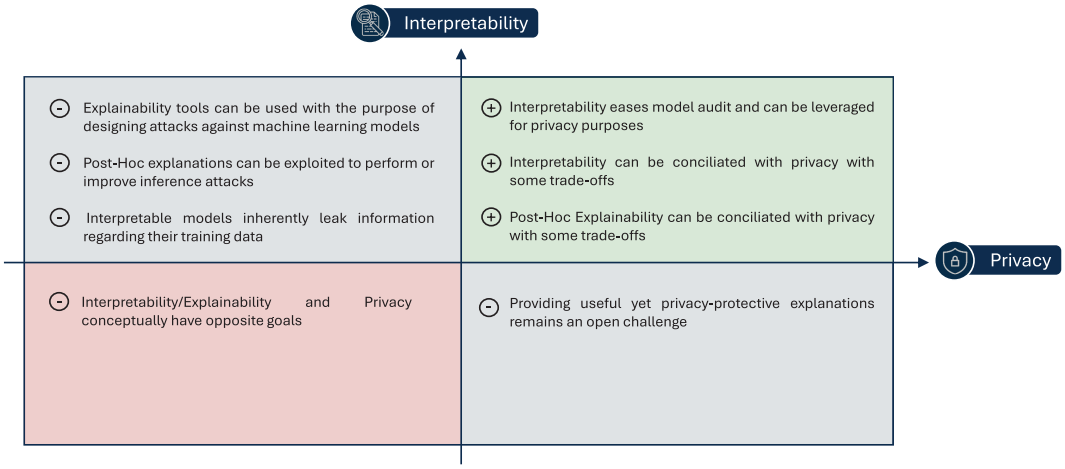

FIGURE 3 | Summary of the identified tensions (−) (Section 5.1) and compatibilities/synergies (+) (Section 5.2) between interpretability and privacy in machine learning. Privacy is represented conceptually along the x-axis: observations on the left tend to harm privacy, while those on the right tend to enhance it. Similarly, interpretability is represented on the y-axis: observations near the bottom tend to harm interpretability or explainability, while those near the top tend to enhance it.

protection, fairness enhancement, or both. This work complements our study well, as federated learning is not our main focus, yet it introduces its own set of constraints, concepts and challenges. Finally, a recent thesis [16] focuses on the interactions between transparency and fairness, thus providing a deepening of (part of) our dedicated Section 3 on this topic. We also want to point out that this survey covers work published up to early 2024 and does not aim to be exhaustive with respect to more recent developments in this rapidly evolving field, although a few of them are included.

The outline of the paper is as follows. First in Section 2, we review the background regarding the three considered aspects of responsible ML, namely fairness, interpretability, and privacy before surveying their interplays. More precisely, Section 3 considers both fairness and interpretability, Section 4 studies the interactions between fairness and privacy, and Section 5 summarizes the connections between interpretability and privacy. Then, Section 6 concludes with the identified remaining key challenges. Finally, Figures 1–3 graphically summarize all analyzed interplays, showing how each desideratum positively or negatively influences each other.

2 | Background

In this section, we first briefly overview the considered machine learning setup and notations. Then, we introduce the three identified pillars of responsible machine learning. For each of them, we briefly review their key ideas, with an emphasis on the particular aspects that will ease the understanding of subsequent sections.

2.1 | Considered Machine Learning Setup

The high-level objective of a machine learning algorithm $\mathcal{L}$ is to leverage a training dataset $\mathcal{D}$ to build a predictive model $\mathcal{L}(\mathcal{D}) = h$. The dataset $\mathcal{D}$ consists of multiple examples (also referred to as samples or observations), each described by a set of continuous or discrete attributes (also called features). In the supervised learning setting—the main focus of our survey—each example is associated with a label, which is discrete in classification tasks and real-valued in regression problems. The objective of the learning algorithm is to construct h such that it can accurately predict the label of previously unseen examples. The proportion of examples, from a separate test dataset disjoint from $\mathcal{D}$ used to assess the generalization ability of h, for which h correctly predicts the label is referred to as accuracy, and serves as a proxy for the model's utility.

While additional desiderata—introduced in the remainder of this section—must also be considered, maximizing accuracy remains the core objective in most machine learning contexts. More fundamentally, it is precisely the pursuit of models that closely capture the patterns in the training data that gives rise to complex tensions with other ethical considerations. Indeed, a trivial model (e.g., one that outputs a constant value) can easily satisfy perfect fairness (by assigning identical predictions to all examples), interpretability (due to its simplicity), and privacy (by ignoring the training data entirely), but does not provide any utility.

2.2 | Fairness

Different approaches to fairness have been proposed in the literature, which can be grouped into three main categories [17]. The rationale of *statistical fairness*, also coined *group fairness*, is to ensure that a given statistical measure has similar values between several *subgroups*, defined by the value(s) of some sensitive feature(s). For example, the statistical parity fairness metric aims at equalizing the positive prediction rate across the different groups, while the equal opportunity metric considers the groups' true positive rates and finally the equalized odds metric handles both their true positive and true negative rates. The underlying principle is that such sensitive features (e.g., race, gender, etc.) should not influence the predictions. *Individual fairness* approaches build on the idea that similar individuals should be treated similarly [18]. For instance, this can be formulated as a Lipschitz condition over the classification function, in which bounding the distance between two examples also bound the distance between their outputs from the model. *Causal fairness* techniques analyze the causal relationships between sensitive features, non-sensitive ones and the target decision, leveraging causal graphs [19].

Depending on which step of the (supervised) ML pipeline they intervene on, fairness-enhancing methods can be divided into three main categories [20–22]. *Pre-processing* methods aim at removing undesired correlations from the training dataset $\mathcal{D}$ before applying standard learning techniques on the sanitized data while *post-processing* techniques modify the outputs of a trained model h to achieve fairness. Finally, *in-processing* (also called *algorithmic modification*) techniques directly adapt the learning procedure $\mathcal{L}$ to produce inherently fair models.

2.3 | Explainability/Interpretability

There are two main approaches toward facilitating the understanding of ML models [23]. On the one hand, post hoc explanations [6] can be crafted to explain the behavior of a black-box model h. Note that h is typically considered a black box when its internal parameters are not accessible—for instance, if it is only available via an API—or when it is so complex that its predictions or underlying rationale cannot be understood by humans. Depending on their form, different types of post hoc explanations can be defined, among which *example-based explanations* consist of data points, belonging to the same space as the model's training set examples. For instance, they can be highly influential training examples [24], nearest neighbors or prototypes. *Counterfactual explanations* also fall into this category, as they are datapoints close to the explained instance but exhibiting a different prediction from the considered model. *Feature-based explanations* take the form of a vector in the feature space, in which each coordinate is the degree to which the associated feature influences a model's prediction. For example, in computer vision, saliency maps [25] highlight the regions of an input image that most contributed to the model's decision. Feature-based explanations can be computed using several mechanisms. For instance, *gradient-based* methods compute the gradients of a model (e.g., a deep neural network) with respect to the input features, either for a given class or for intermediate component(s) of the network, which enables to determine which features contribute the most to a particular prediction. In contrast, *perturbation-based* methods modify the input provided to the black-box and observe the resulting changes in the model's outputs.

On the other hand, one can learn models h that are inherently interpretable by humans, by considering appropriate learning algorithms $\mathcal{L}$. For instance, decision trees or rule lists of reasonable size are commonly considered as interpretable [26]. While the meaning of a *reasonable size* is ill-defined and context-specific, it indicates that model simplicity is a crucial property to consider while building these models. *Sparsity*—which quantifies the size of the learned model (e.g., in terms of depth, number of nodes or number of rules)—is often used as a proxy for model simplicity.

2.4 | Privacy

The development of privacy-preserving mechanisms for ML has been widely motivated by the flourishing literature on inference attacks against models in recent years. In the generic setting, such attacks leverage the outputs of a computation to retrieve information regarding its inputs [27]. More specifically in machine learning, the computation typically consists of a learning algorithm $\mathcal{L}$, which takes as input a training dataset $\mathcal{D}$ and produces a trained model h as output. Two distinct adversarial settings are generally considered in the literature. In the *black-box* setting, the adversary does not know the model's parameters and can only query it through an API. In contrast, in the *white-box* setting, the adversary has full knowledge of the model parameters. Of course between these two extreme scenarios, diverse *gray-box* settings are possible.

Different types of inference attacks have been proposed against ML models [4, 28], among which:

- *Membership inference attacks* try to infer whether given examples were used to train a model or not [29].
- *Reconstruction attacks* aim at reconstructing part of a model's training data [27].
- *Model extraction attacks* aim at stealing a black-box model's internal functionalities or parameters [30].
- *Model inversion attacks* focus on retrieving a model's inputs by only observing the associated outputs [31]. Hence, such attacks often target the data provided at inference time (and not solely the training data).

To counter these risks, several syntactic models of anonymity were proposed. More precisely, these approaches consist in grouping examples within *blocks* so that the profile of a user is indistinguishable among those belonging to the same block [32]. For instance, k-anonymity [33, 34], requires that each block contains at least k examples. Several extensions of k-anonymity were proposed, among which ℓ-diversity requires that at least ℓ different values of the private features are *well represented* within each block, and t-closeness [35] additionally ensures that the distribution of these values is sufficiently close to that of the entire dataset.

Nonetheless they are not well-adapted to ML and do not provide formal privacy guarantees. Thus, *differential privacy* (DP) has been adopted as the leading approach, in parts because it can be used to precisely bound the amount of information the output of a computation leaks regarding its inputs [36]. Due to the strong theoretical guarantees it provides, to the interesting properties it exhibits, and to the availability of several mechanisms to enforce it, it has now been widely adopted. Examples of recent applications of DP include the 2020 release of the U.S. Census Bureau[3] [37], but also its use by companies such as Google [38], Facebook [39] and Apple [40].

Referring to (ϵ, δ)-DP, two parameters help control the level of enforced privacy. Intuitively, ϵ bounds the contribution of each individual example to the output of the computation, while δ corresponds to the probability of privacy failure, with tighter values of these parameters indicating a stronger privacy protection. *Pure DP* refers to scenarios in which $\delta = 0$ while *approximate DP* covers cases in which $\delta > 0$. DP exhibits several important properties, among which the immunity to post-processing, which states that the output of a differentially private algorithm remains differentially private whatever (data-independent) computations are further performed on it. Several mechanisms were proposed to enforce DP [41]. For instance, the *Laplace* (respectively, *Gaussian*) *mechanism* [36] adds random noise drawn from a Laplace (respectively, Gaussian) distribution to the computed value, with the noise magnitude being scaled to the function's sensitivity (i.e., the maximum impact a single individual can have on the computation's output). The *functional mechanism* [42] approximates the function using its polynomial Taylor expansion and perturbs the coefficients of the resulting polynomial form with noise. Unlike the aforementioned noise addition techniques, the *exponential mechanism* [43] consists in drawing an output from a probability distribution, in which the probability of a candidate depends on its utility.

Several frameworks for differentially private ML exist [44, 45], which typically incorporate the DP mechanisms into an existing learning algorithm $\mathcal{L}$ to ensure that the resulting model h satisfies DP guarantees. For instance, `DP-SGD` [46] was proposed to train deep learning models under DP. The authors have modified the traditional Stochastic Gradient Descent (SGD) by clipping the norm of the computed individual gradients (to bound each example's contribution to the computation) before perturbating them with Gaussian noise. Another approach based on ensemble methods, called `PATE`, considers a particular setup, with a private training set and a public unlabeled one [47, 48]. First, the (private) training set is partitioned into a number of non-overlapping subsets used to train a set of *teacher* models. Afterward, the predictions of the teachers (i.e., vote histograms) are made differentially private by adding Laplace noise. The public data is then labeled using these noisy predictions, and used to train a differentially private *student* model. We refer the interested readers to the recent survey of [49], which reviews existing techniques to make supervised learning algorithms differentially private.

3 | Fairness and Interpretability

In this section, we first review the tensions between fairness and interpretability before exploring some synergies.

3.1 | Tensions

First, we elaborate on the theoretical and empirical tensions between fairness and simplicity, which is often considered as a proxy for interpretability. Afterward, we discuss the main challenges that need to be tackled when jointly pursuing the interpretability and fairness desiderata. Finally, we list different ways in which post hoc explanations can be unfair.

3.1.1 | Tensions Between Fairness and Simplicity

3.1.1.1 | Simplicity and Fairness Intrinsically Conflict. A framework to theoretically study the implications of enforcing interpretability is proposed by [50], adapted from that of [51]. It considers simplicity as a proxy for interpretability. More precisely, a ML model is represented as a set of cells partitioning the input space and simplifying a model consists in merging some of its cells (hence diminishing their number and the model's complexity). The authors prove that, for every non-trivial group-agnostic simplification, there exists a more complex classifier that simultaneously strictly improves both accuracy and (statistical) fairness. This classifier can be efficiently constructed by carefully selecting some examples from chosen subgroups and splitting their associated cells. Overall, this result suggests that interpretability/simplicity comes at some cost in terms of accuracy/fairness. Similar results were originally shown by [51], further illustrating how simplicity can be inconsistent with statistical fairness notions. As stated by [52], while model interpretability is an abstract notion, enforcing it can only reduce the set of admissible ML models. Consequently, ensuring interpretability can only decrease the (training) accuracy. A similar reasoning can also

be done with respect to fairness. More precisely, by limiting the space of admissible classifiers, the enforcement of fairness reduces the number of possible trade-offs, which can be an obstacle to achieve both fair and accurate learning.

A handful of recent studies have sought to precisely characterize these trade-offs. In particular [53], have derived theoretical bounds on the statistical fairness levels achievable by any predictive model on a given dataset and task, while maintaining accuracy within a bounded margin from the best-performing model. Similarly [54], provide tight statistical fairness bounds tailored to specific hypothesis classes and sparsity levels, again under constraints on the acceptable drop in accuracy. Their empirical results with two types of interpretable models (namely, scoring systems and decision diagrams) allow a precise (and certifiable) quantification of the optimal trade-offs between predictive performance, statistical fairness (statistical parity and equal opportunity), and sparsity (respectively, number of non-zero coefficients and number of active nodes). More precisely, they have shown that imposing stringent sparsity requirements can disproportionately affect—and even systematically discriminate—minority groups.

3.1.1.2 | Empirical Trade-Offs Are Complex. An empirical study of the trade-offs between interpretability and fairness was conducted by [55]. In this study, the number of features available to a classifier is used as a measure of its complexity and acts as a proxy for interpretability. By changing this number, the authors report the variations obtained with respect to statistical fairness notions (namely, statistical parity and equal opportunity). The experiments on synthetic and real-world datasets show several trends that mainly depend on the correlation between sensitive attributes, non-sensitive ones as well as class labels. As expected, when the sensitive attribute is correlated (even moderately) with the class label, using it explicitly greatly increases the model's unfairness. The results obtained rely strongly on the chosen notion of interpretability and as such cannot be considered generic. In addition, they demonstrate that the trade-off between fairness and interpretability is, in practice, complex and data-dependent. In a later work [56], propose the notion of *decision complexity*, which is defined as "the minimum number of parameters needed for the classifier to make a prediction on a new data point". By generalizing the concept of sparsity, this metric enables quantitative comparisons of interpretability across different types of predictive models. It can also be used to empirically assess the trade-offs between accuracy, fairness and interpretability achieved by several state-of-the-art fair learning algorithms. The experiments conducted reveal a so-called *price of interpretability*, as more interpretable models often result in lower performance for a fixed fairness level. However, the observed trends depend strongly on the specific fair learning algorithm used.

3.1.2 | Combining Fairness and Interpretability is Challenging

3.1.2.1 | Learning Optimal Interpretable Models Under Fairness Constraints is Computationally Challenging. Due to their combinatorial nature, learning optimal interpretable machine learning models under constraints (e.g., fairness constraints) has been identified as one of the main technical challenges toward interpretable machine learning [57]. While approaches producing optimal interpretable and fair ML models exist in the literature (e.g., an Integer Programming formulation for learning optimal fair decision trees), they are often computationally expensive and difficultly scale. Yet, recent work shows that the conflict between accuracy and fairness can be leveraged to perform an effective pruning (using Integer Linear Programming) when learning optimal fair rule lists [58].

3.1.2.2 | Explanations May Not Preserve Fairness Properties of a Model. It was observed by [59] that popular explainability frameworks may not reliably reflect the fairness properties of the explained models. For example, on the one hand it is possible to compute post hoc explanations that appear to be fair to explain an unfair black-box model [60]. On the other hand, the explanations of a fair model's decisions may (wrongly) rely on sensitive features and exhibit discrimination [61]. In addition, the choice of the explanation method as well as the type of explanation it produces both impact the users' perceived fairness [62]. The fairness of post hoc explanations generated from a fair model's decisions was also investigated by [59]. More precisely, based on group fairness notions, the fairness of an explanation can be formulated similarly to that of a classifier (an explanation being seen as a local surrogate model). Afterward, fairness is computed on a neighborhood of the explained example. For such artificial points, no label is known, which means that only the statistical parity metric can be used. These researchers show that the fairness property of the explained model may not be reflected in the generated explanations and propose a framework for producing fairness-preserving explanations.

3.1.2.3 | Fairness-Enhancing Methods May Require Non-Interpretable Transformations, Hence Harming Interpretability. In a study on interpretable, fair and accurate ML for criminal recidivism prediction [63], observe that fairness-enhancing methods often require non-interpretable transformations, which are not compatible with interpretability desiderata. Indeed, pre-processing methods usually perform complex transformations of the input features, which harm their original semantic [64, 65]. The resulting representation hence cannot be used to produce an understandable model. Furthermore, the corrections performed to a model's outputs by post-processing techniques [66] can also lead to non-interpretable processes.

3.1.3 | Other Unfair Effects of Explainability Methods

3.1.3.1 | Post Hoc Explanations Affect Individuals' Privacy in a Disparate Manner. As discussed later in Section 4.1, minority groups often suffer from increased privacy risks. Interpretability can also exhibit this trend, as noted by [67, 68]. For instance, when investigating whether membership information can be inferred from post hoc explanations, it has been observed that outliers as well as "hard to generalize" examples belonging to minority groups are at a higher risk of being disclosed than majority groups. This is partly due to the fact that they are more susceptible of being part of the generated explanations. In

such case, interpretability tools can penalize minorities by leaking more information about disadvantaged groups.

3.1.3.2 | Post Hoc Explanation Frameworks Can Introduce Unfairness Through Disparity in Explanation Quality. Group-based disparities in explanation quality have been recently investigated by [69]. More precisely, the authors first identify key characteristics that define the quality of an explanation (e.g., fidelity, stability, consistency, and sparsity). Then, they conduct a large experimental study demonstrating that there is often a disparity in the quality of the explanations produced affecting minority groups. Such quantitative disparity is identified to depend on the type of model being explained and on the particular post hoc explanation framework considered. Using several real-world applications (e.g., finance, healthcare, college admissions, and the US justice system) and post hoc explanation frameworks [70], have also demonstrated that the fidelity of the produced explanations varies significantly across the different identified subgroups of the population. Finally, they suggest that robustness techniques can help reduce the observed disparity—but emphasize that communicating details regarding such disparity to end-users is critical.

3.1.3.3 | Counterfactual Explanation Frameworks Can Harm Subgroups of the Population by Consistently Providing Higher Cost Recourse. In the context of counterfactual explanations, the *cost of recourse* is defined as the amount of effort a user has to do to implement the provided recourse and change the model's decisions. In this context, it was shown that counterfactual explanation frameworks may provide lower cost recourse for some subgroups of the population while harming some others [71, 72]. For instance, some minority groups may have to make more effort to implement the provided recourse after a loan refusal. To face this issue, *recourse fairness* was studied [73, 74] and frameworks equalizing the cost of recourse across subgroups were proposed.

3.1.3.4 | Post Hoc Explanations Can Be Manipulated. Explainability tools are designed to facilitate model audit and enhance the users' understanding. However, because the process of explanation generation can sometimes be opaque, post hoc explanations could potentially be manipulated by black-box model holder to hide unfair decision-making processes by providing manipulated fair explanations. Indeed, it was shown that black-box explanations can be misleading, for instance by achieving high fidelity with respect to the explained model while using entirely different features, leveraging correlations in the feature space [75]. In addition, it has been demonstrated that this can be exploited and extended to an existing framework [76] to generate explanations favoring some given features while avoiding others. Finally, the authors have conducted a user study and find out that misleading explanations can increase the user trust in black-box models wrongly.

Other works have also shown how malicious entities can manipulate explainability techniques to hide the true reasoning of the underlying model. For example, it is possible to directly craft manipulated explanations, such as local surrogate models [60, 77] that appear fair but actually explain the output of a globally unfair black-box, with such practice being coined as "fairwashing". Explanation frameworks can also be potentially manipulated, for instance by detecting artificial examples generated by perturbation-based methods and giving them a chosen output value [78]. This can be leveraged to hide a black-box model's unfairness by crafting and providing fair explanations to an auditor [79]. Furthermore [80] and [81], have shown that it is possible to fine-tune a pre-trained model to manipulate the output of feature importance explanation methods while having little impact on the model's accuracy. Considering sequence classification and sequence-to-sequence tasks (i.e., in which the input to the model is a sequence of words) [82], propose a method to train a model with significantly reduced attention mass over some chosen words (e.g., gender-related prefixes) while still using them for prediction. A user study shows that the proposed method is able to mislead users into thinking that the underlying model is fair, while it is actually biased against gender.

It was also shown to be possible to learn a model so that the counterfactual explanations generated by some off-the-shelf algorithm look *recourse fair* across subgroups of the population (i.e., the cost of the recourse associated to the counterfactual explanations does not vary too much between individuals from the different subgroups), while also being able to generate lower cost recourse explanations for some privileged subgroup(s) by simply adding a small adversarial perturbation [79, 83]. In [84]'s work, an adversary is able to generate adversarial examples with chosen prediction by a black-box model that also fool popular explainability tools. This illustrates the fact that post hoc explainability techniques are not a reliable way to detect adversarial inputs manipulation. Finally [85], consider the setup of a fairness audit in which the data is private and owned solely by the malicious model holder, which provides subsamples to the external auditor. They show that the former can manipulate the auditor's explainability methods to hide unfair decision-making (such as the influence of a sensitive attribute) by providing adversarially selected data samples. In addition, such practices are particularly difficult to detect in a remote setting, in which the explanation is provided by a third-party API [86].

Finally, although many tensions between explainability/interpretability and fairness exist, one can still identify some synergies, as discussed hereafter.

3.2 | Synergies

3.2.1 | Interpretability and Explainability Ease Model Audit

As mentioned by [87], it is easier to detect and debate possible biases or unfairness issues with an interpretable model than with a black-box one. This inherent benefit of interpretable models applies both to fairness and accuracy, as it makes it possible to detect and correct possible inaccuracies with respect to the training data—which is more difficult with black-box models. Following the same line of research [88], state that interpretability can be used to qualitatively ascertain whether other desiderata—such as fairness—are met. Post hoc explainability methods can also facilitate fairness audit by gaining insight regarding the causes of a model's unfairness. For instance [89], propose to rely on *fairness explanations* based on Shapley values to be able to attribute a model's overall unfairness to individual input features.

3.2.2 | Fairness Can Act as a Regularizer

It was observed in the literature that enforcing fairness constraints can have a regularizing effect, thus also reducing overfitting [90]. More precisely by preventing over-complex models, this can lead to sparser and more interpretable models.

4 | Fairness and Privacy

In this section, we first highlight the identified theoretical and empirical tensions between fairness and privacy. We then review some synergies illustrating how the two requirements can be conciliated. Note that part of this intersection is covered in much more details by a recent survey [12] studying the interactions between fairness and differential privacy (DP), in both decision making and machine learning tasks.

4.1 | Tensions

As discussed in Section 2.2, it is desirable and often legally required to ensure that sensitive attributes do not directly or indirectly influence the predictions of a ML model. However, while many popular fairness-enhancing approaches require the availability of such sensitive attributes, their collection and use may be prohibited by privacy regulations or anti-discrimination laws. Some approaches propose to use an encrypted version of the sensitive attributes so that the users do not have to explicitly reveal this information. For instance [90], leverage cryptographic approaches such as Secure Multi-Party Computation (SMPC) to build a fair model. Nevertheless, processing encrypted information ensures that the computation does not leak anything more than its outputs, but does not protect them from inference attacks. This illustrates a first, straightforward intrinsic conflict between fairness and privacy. Furthermore, when applied jointly, both notions often conflict, as discussed in more details in the following paragraphs.

4.1.1 | Group Fairness and Differential Privacy Are Theoretically Incompatible

It is provably impossible to build ML models strictly respecting a given group fairness constraint while respecting DP. More precisely [91], have shown that pure $(\epsilon, 0)$-DP and fairness (more precisely equal opportunity) cannot be simultaneously satisfied without reaching trivial accuracy. The authors have noted that this holds for pure $(\epsilon, 0)$-DP, but is also applicable for approximate (ϵ, δ)-DP (as δ is usually required to be cryptographically small). An impossibility theorem is also stated by [92], considering popular group fairness definitions: *if a learning algorithm* $\mathcal{L}$ is $(\epsilon, 0)$*-differentially private and is guaranteed to output an approximately fair classifier, then* $\mathcal{L}$ *is constrained to output a constant classifier*. The idea of the proof is essentially the same as that of [91]. (i) Consider a learning algorithm $\mathcal{L}$ satisfying pure $(\epsilon, 0)$-DP, for any two datasets $\mathcal{D}$ and $\mathcal{D}'$, and for any classifier h, if $\mathcal{L}$ outputs h for $\mathcal{D}$ with probability strictly greater than zero, then it must output h for $\mathcal{D}'$ with strictly positive probability too. This can be proved because, for any two datasets $\mathcal{D}$ and $\mathcal{D}'$, it is possible to build a series of datasets neighboring two-by-two, from $\mathcal{D}$ to $\mathcal{D}'$ (and the property must be verified for all pairs of neighboring datasets by definition of pure DP). (ii) Recall that $\mathcal{L}$ can only output classifiers respecting a given (exact or approximate) fairness requirement: if a classifier h does not meet the fairness requirement on the training set $\mathcal{D}$, then $P(\mathcal{L}(\mathcal{D}) = h) = 0$. The conjunction of (i) and (ii) implies that $\mathcal{L}$ can only release constant classifiers (and hence pure DP and group fairness cannot be satisfied jointly).

4.1.2 | Enforcing Fairness Increases Privacy Vulnerabilities

Disparities with respect to the vulnerability to Membership Inference Attacks (MIAs) between various subgroups of the population are observed by [93]. The theoretical analysis suggests that vulnerability to MIA is caused by *distributional overfitting*, which quantifies the distance between the distributions of outputs of the model on the training set and outside. Disparate vulnerability to MIAs arises if and only if distributional overfitting differs across subgroups. In practice, as aforementioned in Section 3.1.3, subgroups that are inherently more difficult to fit and/or that are less represented in the data are indeed more vulnerable to MIAs. Additionally, overfitting can increase these vulnerabilities, but also their disparities. For instance, it was empirically shown that enforcing fairness constraints may help under certain conditions, but can also exacerbate the observed disparities or even create new ones in real-world applications. Finally, the authors have recalled that DP[4] upper bounds the vulnerability of all individuals or subgroups, hence also upper-bound their disparity. However, it does not remove it completely and in addition to get an interesting mitigation, the privacy budget must often be really tight, hence resulting in utility drops.

In a position paper [94], emphasize the importance for a privacy-preserving mechanism to protect individuals with equivalent effectiveness. However, while DP4 provides the same (worst-case) theoretical protection for all dataset examples, the actual privacy vulnerability is often not uniformly distributed. The privacy implications of fairness are empirically studied by [95], quantifying the *data privacy risk* as the success of a black-box MIA. The authors have empirically shown that enforcing fairness constraints disproportionately raises the privacy risk of the unprivileged subgroups: "fairness comes at the cost of privacy, and the privacy cost is not equal across subgroups". This is explained by the fact that the fairness requirements they have used requires the model to equally fit the unprivileged subgroups. When such subgroups are smaller, each example has a stronger impact over the resulting model and, in the worst case, is memorized. In addition, the more unfair the unconstrained model is, the higher the privacy vulnerability disparity will be, as there is more unfairness to be compensated.

Finally, information regarding a model's fairness can be exploited to reconstruct the sensitive attributes of its training examples [96, 97]. These works rely on declarative programming approaches to encode the fairness desiderata and perform (or improve) the reconstruction. Their empirical results demonstrate that the information brought by fairness regarding sensitive attributes can effectively by exploited by an adversary to harm the privacy of individuals involved in the model's training data.

4.1.3 | Differential Privacy Disproportionately Affects Utility

The effects of enforcing differential privacy on a model's accuracy on different subgroups of the population are studied by [98], using the *accuracy parity* fairness notion, which equalizes the model's accuracy across the subgroups. Considering several image classification and natural language tasks, they use the popular DP-SGD [46] framework for differentially private deep learning in both centralized and federated settings. This large empirical study shows that gradient clipping and random noise addition, the key mechanisms of DP-SGD, disproportionately affect underrepresented subgroups. Indeed, enforcing (approximate) DP leads to higher accuracy drops for minorities and discriminated groups, such as darker-skinned people in the context of facial recognition, but also at the intersections of different subgroups. This leads to a "poor gets poorer effect", in which the classes with low accuracy in the non-DP setting suffer the largest accuracy drops when applying DP. In a follow-up work [99], empirically observe that the differentially private PATE [47, 48] framework (introduced in Section 2.4) also has disparate impact on the resulting model's utility. However, they report that PATE has smaller disparate impact compared to DP-SGD to reach similar approximate DP levels, and note that a sweet spot for the number of teachers exists, which minimizes the induced disparities [100] observe that the accuracy disparity caused by (approximate) DP still occurs even when the data is slightly imbalanced, and for loose privacy guarantees. Indeed, two main factors were identified in the literature to explain this effect: properties of the training data, and characteristics of the DP mechanism, which are summarized and analyzed with more details in a recent survey [12].

It was also observed in healthcare applications (x-ray images classification and mortality prediction in time series) that small groups and samples at the tail of the data distribution suffer from a larger accuracy drop compared to majority groups and typical examples [101]. Furthermore, the characteristics of DP learning mechanisms themselves are also directly related to the magnitude of the observed disparate impact. This encompasses the gradient clipping and noise addition mechanisms of DP-SGD (as aforementioned), as well as the size of the teacher ensemble and the confidence of the voting teachers in PATE [102]. Different technical solutions to mitigate the disparate impact of DP on a model's utility were proposed. Indeed, it was shown that it is possible to modify DP-SGD to use different clipping bounds for the different identified subgroups [103]. Other work [104] performs early stopping based on a public validation set. When using PATE in low voting confidence regimes, small perturbations may significantly affect the result of the voting result. To mitigate this phenomenon [102], propose to use soft labels and report confidence scores associated with each target label, rather than reporting solely the label with the largest confidence. While being heuristic as it does not guarantee any form of fairness, these approaches have been empirically shown to reduce the disparate impact caused by traditional DP mechanisms.

The disparate impact of DP mechanisms was also observed for decision tasks [11] have studied the setup in which agencies release differentially private versions of their databases that are then used for several allocation problems. The authors consider three real-life allocation problems using the differentially private Census data: namely printing of election materials in minority languages, allocation of funds to school districts to assist disadvantaged children and apportionment of legislative representatives. They demonstrated that the noise added by (pure) DP mechanisms leads to errors in the computed allocations compared to the true allocations (i.e., the allocations that would be decided without DP). The key point of their work is that this error affects the entities being allocated some resources in a disparate manner. For instance, it is empirically shown that small school districts often benefit an overestimated allocation. On the other side, larger district may get a smaller allocation, which harms their enrolled children. This effect was also observed in the literature with two main causes being identified [12]. In a nutshell, the shape of the decision problem can disproportionately exacerbate the noise added by the DP data release if it involves non-linearities in its computation, such as thresholds for funds allocation. Additionally, post-processing steps can induce intrinsic biases. For instance, ensuring simple non-negativity constraints within the computed values can imply a positive bias. It was also shown that DP mechanisms adding data-dependent noise are responsible for a more important disparity, due to the fact that, contrary to standard DP mechanisms (such as the Laplace mechanism), the effect of DP differs between entities. Finally, other aspects of privacy can also impact fairness. For instance, recent work by [105] show that models designed to take into account potential future unlearning requests, which are request in which a user asks for the contribution of his data to be removed from the model, disproportionately affects the utility for minority groups.

4.1.4 | Differential Privacy Disproportionately Affects the Quality of Post Hoc Explanations

Datta et al. [106] propose the notion of differentially private post hoc explanations, among which some aim at identifying proxy features that cause a *group disparity* (i.e., a difference in the average prediction between several subgroups). Then, it is shown that, for minority groups, the amount of noise required to make the explanations differentially private results in a significant loss in its utility, hence making more difficult the discovery of discriminatory proxy features. While proposing a framework to generate differentially private post hoc explanations [107], have observed that sparse data regions, which often correspond to underrepresented subgroups are associated to poorer performances, either in terms of required privacy budget or explanation quality. In both cases, privacy disproportionately affects minority groups, which is consistent with previously mentioned works.

Overall, DP and statistical fairness are both theoretically incompatible and strongly conflict in practice. On the one side, to ensure fairness minority groups, the corresponding examples shall yield a higher importance in the learning process, which exposes their information more than for examples of the majority group. On the other side, to ensure DP, one must reduce more the influence of underrepresented subgroups, as learning an equivalent amount of information for them would result in an increased per-example privacy risk. Nevertheless, in the next subsection, we show that the two notions can be jointly applied under certain

circumstances, and thus that there are some synergies between privacy and fairness.

4.2 | Synergies

4.2.1 | Differential Privacy and Approximate Fairness Can Be Jointly Enforced With Some Trade-Offs

As discussed in Section 4.1, it is impossible for a learning algorithm to satisfy DP while also producing a model strictly complying with fairness constraints. However, it is possible for a DP (pure or approximate) learning algorithm to output a model *approximately* satisfying given fairness criteria [91]. This leads to a trade-off between the DP guarantees and the observed model's fairness. Hereafter, we first introduce different methods of the literature jointly handling differential privacy and fairness.

The notion of Private and Approximately Fair Agnostic PAC (Probably Approximately Correct) Learning was introduced by [91]. It states that a learning algorithm satisfies DP while returning an accurate and approximately fair classifier with high probability. The authors implement this notion using the Exponential Mechanism, with a utility function being the sum of a model's error and unfairness. The sensitivity of the utility function being data-dependent, the Laplace mechanism is used to upper-bound it in a differentially private manner. This approach achieves the desiderata of privacy, fairness, and accuracy, but the running time of the Exponential Mechanism scales linearly with the hypothesis class size, which is exponential for common hypothesis classes. This motivates the need for an efficient algorithm conciliating these desiderata. To realize this, the authors have built upon a polynomial-time algorithm from the literature, producing approximately fair and accurate randomized classifiers with high probability. In a nutshell, this algorithm formulates the fair learning problem as a two-player zero-sum game, between a Learner minimizing error while satisfying fairness constraints and an Auditor updating Lagrangian multipliers to penalize the largest subgroup-wise fairness violations. This algorithm is modified to satisfy DP by using a differentially private subroutine to privately compute the players' best responses in each round.

Two methods are proposed by [108] to achieve jointly DP and fairness in logistic regression. *Decision boundary fairness* is used as a notion of fairness that provably minimizes statistical parity violation. A first approach coined `PFLR` considers the fairness constraint as a penalty term to the objective function. DP is enforced using the functional mechanism [42]. More precisely, the objective function is approximated through its polynomial representation based on Taylor expansion before being perturbed by injecting Laplace noise into its polynomial coefficients. Minimizing the perturbed objective function leads to the computation of differentially private model parameters. A second approach, named `PFLR*` and based on the first one, takes advantage of the connection between ways of achieving differential privacy and fairness. More precisely, the authors noted that adding the fairness penalty is equivalent to shifting the value of some coefficients of the polynomial form of the objective function. Thus, they do not incorporate the fairness penalty term directly in the objective function and rather integrate it via mean-shifting the Laplace noise added to a subset of the coefficients. As such shift is dataset-dependent, a small part of the privacy budget is used to estimate it in a differentially private manner. The Theoretical analysis as well as empirical evaluation show that `PFLR*`, by separating privacy budgets on objective function and fairness constraint, offers a more flexible framework to find good trade-offs among privacy, fairness, and utility.

In a follow-up work [109], extended `PFLR` by proposing to have two distinct privacy budgets to add Laplace noise with larger magnitude to the coefficients of the terms involving the sensitive attributes than to the others within the objective function. They also propose a second approach using the relaxed functional mechanism to enforce approximate DP (ϵ, δ)-DP to improve on utility. It utilizes the extended Gaussian mechanism to perturb the objective, adding random Gaussian noise to the coefficients of the polynomial form of the objective function. Empirical evaluation on real-world datasets confirms that the use of (ϵ, δ)-DP leads to an improved utility in all scenarios compared to pure DP. Furthermore, the use of two distinct privacy budgets can help enforcing stronger privacy guarantees while also reducing the correlations with the sensitive attribute, thus also improving fairness.

A differentially private framework to train deep learning models that satisfy several popular group fairness notions was proposed by [110]. This approach considers the Lagrangian relaxation of the fairness-constrained learning problem, and leverages a Lagrangian dual approach to solve it: the fairness violation terms, weighted by Lagrangian multipliers, are directly added to the objective function. Then, the training procedure consists of iteratively repeating two successive steps: primal and dual. The primal update step optimizes the model parameters to minimize the objective function, given the current Lagrangian multipliers. Afterward, the dual update step updates the value of the Lagrangian multiplier to approximate the stronger Lagrangian relaxation. To enforce differential privacy for sensitive attribute information, differential privacy is achieved at both steps, when computing the fairness violation terms or their gradients. In the primal update step, clipped and noisy gradients are used. The model parameters optimization is done on this noisy version of the objective function (in which only the fairness violation term, accessing subgroup membership which we want to protect, is impacted by the DP mechanism). A similar mechanism is done on the dual update step, in which constraint violations are clipped and perturbed with carefully calibrated Gaussian noise. Extensive empirical evaluation shows that the fairness violation decreases as the privacy budget increases: thus enforcing DP leads to violating more fairness. This is explained by the fact that relaxing the DP constraint allows either to perform more iterations (hence propagating more fairness violation information) or to inject less noise for a fixed number of iterations (hence propagating more accurate fairness violation information). Another surprising trend is that the model accuracy slightly decreases as ϵ increases. This is due to the fact that enforcing weaker DP allows the fairness constraints to have more impact on the objective function, hence penalizing more the accuracy.

Two fair learning algorithms have been adapted by [111] to satisfy both fairness (here in terms of equalized odds) and DP (with respect to the sensitive attributes). They first consider the

post-processing method of [112]. In a nutshell, given a pre-trained and possibly unfair classifier, the approach first computes its per-group per-ground truth prediction proportions. It then solves a Linear Program to compute per-group per-class prediction probabilities defining a fair randomized classifier. To enforce (pure) ϵ-DP in this setting, the authors simply add well-calibrated noise drawn from the Laplace distribution to the computed statistics before solving the LP with them. Theoretical analysis of how the introduced noise propagates to the solution of the LP leads to bounds on accuracy and fairness violation that are met with high probability. This quantifies a trade-off between accuracy, fairness, and privacy: weaker DP guarantees lead to tighter bounds on accuracy and fairness, while stronger DP guarantees (satisfied by adding more noise) increase the bounds, and the possible loss on accuracy and fairness. Experimental evaluation demonstrates that this simple method is able to provide interesting trade-offs even with small datasets but is expected to perform worst than the second approach on large ones. The later builds upon an in-processing approach [113], which formulates the problem of learning a fair and accurate classifier as finding the equilibrium of a two-player min-max game. A Learner minimizes the objective function over the set of possible classifiers while an Auditor maximizes it by choosing the value of the multipliers penalizing fairness violations. To enforce (approximate) (ϵ, δ)-DP, the authors add well-calibrated Laplace noise while computing the gradients of the Auditor, and use the exponential mechanism for the Learner's model selection. Similar to the first case, a stronger privacy guarantee (smaller ϵ and δ) leads to weaker accuracy and fairness guarantees. However, a new trade-off can be controlled through the maximum norm of the multipliers: larger values lead to tighter fairness bounds but looser error bounds, and vice-versa. For both approaches, introducing noise to achieve DP leads to a reduction in the fairness guarantees (similarly to accuracy).

Mozannar et al. [114] consider the setup in which the sensitive attributes are released using local DP (i.e., a variant of DP in which each user locally randomizes his data before releasing it), and propose a two-step approach. First, a classifier that is fair with respect to the noisy sensitive attributes is built, using a state-of-the-art in-processing fair learning algorithm [113]. Second, a modified version of a post-processing fairness-enhancing method [112] is used to ensure with high probability that the model is also fair with respect to the (unknown) original sensitive attributes. For strong privacy regimes, this post-processing step is empirically shown to significantly decrease the fairness violation [115] have studied a similar local DP setting, extending it to scenarios involving multiple sensitive attributes. More precisely, they have introduced a privacy budget allocation strategy that adjusts to the domain size of each sensitive attribute, ensuring a more balanced application of noise. Their extensive empirical evaluation shows that applying local DP to sensitive attributes before training can slightly improve fairness across most—though not all—statistical fairness metrics and datasets. This intriguing result suggests that using noisy versions of sensitive attributes may, in some cases, enhance fairness by weakening correlations between sensitive features and labels. However, it also complicates the enforcement of strict fairness constraints, as highlighted in aforementioned studies. This result is consistent with the observations of [116], who found that models trained using `DP-SGD`, combined with proper hyperparameter optimization, often maintain or even slightly improve on statistical fairness metrics—including statistical parity and equalized odds—when compared to non-private baselines.

More recently [117], have proposed `FairDP`, a framework for training deep learning models that simultaneously satisfy differential privacy (DP) and statistical fairness criteria such as statistical parity, equal opportunity and equalized odds. The approach builds on the principles of `DP-SGD`, using gradient clipping and noise addition, but with a key innovation: at each gradient descent step, group-specific noisy model updates are computed independently and then aggregated. This ensures that all demographic subgroups contribute equally to the training process. Crucially, `FairDP` leverages the known distribution of the injected DP noise to derive probabilistic bounds on group fairness metrics. This allows the method to provide formal fairness certifications, despite the inherent randomness introduced by DP. Empirical evaluations show that `FairDP` significantly improves fairness (e.g., over 65% improvement on key metrics) and complies with DP with only a modest reduction in accuracy (less than 4%), outperforming baseline methods across multiple benchmarks.

4.2.2 | The Fairness Cost of Differential Privacy Can Be Theoretically Bounded

Recent work theoretically shows that the impact of (pure or approximate) DP on fairness is bounded and can be computed to obtain non-trivial guarantees regarding the private model's fairness [118]. The underlying analysis relies on the fact that, just like a model's accuracy, common statistical fairness metrics are pointwise Lipschitz continuous with respect to the model parameters. Then, proving that the private model is sufficiently close to the optimal non-private one implies that their fairness is also close. Interestingly, the theoretical bound tightens linearly with respect to the size of the training set: the "loss of fairness" due to privacy vanishes when the number of training examples increases. Previous works also introduced probabilistic bounds on both fairness violations and accuracy loss in differentially private (DP) algorithms. Notably [91], and [111] have derived high-probability bounds that quantify the deviation in fairness and accuracy of their DP algorithms relative to their non-private counterparts. These bounds follow common patterns: they improve (i.e., become tighter) when one increases the privacy budget, enlarges the training dataset or restricts the hypothesis class complexity. Interestingly, the in-processing method of [111] introduces an additional hyperparameter that allows explicit control over the trade-off between accuracy and fairness, in which tightening one bound necessarily loosens the other.

4.2.3 | The Privacy Cost of Fairness Audits Can Be Bounded

Online platforms often use machine learning techniques to perform recommendations or other predictions involving individual's data. Because their outcomes can possibly harm some users, it is necessary to audit their fairness properties. However, this raises important privacy challenges, as the data used to train the models (and its distribution) is often private, and revealing it to (even trusted) third-parties increases the risk of disclosure.

Recent work [119] considers fairness audits of social media algorithms. They propose auditing techniques that come with fairness guarantees and have bounded impact over the privacy risk, which shows that the two concerns can be conciliated with bounded cost over one another.

4.2.4 | Individual Fairness and Differential Privacy Are Both Robustness Definitions

As introduced in Section 2.2, individual fairness can be formulated as a Lipschitz condition: just like DP, it is a robustness definition [120]. More precisely [18], has observed that individual fairness constitutes a generalization of differential privacy. The authors draw an analogy between individuals in the setting of fairness and databases in the setting of differential privacy. Indeed, as also noted by [65], differential privacy requires that "algorithms behave similarly on similar databases", while individual fairness enforces that classifiers yield similar outcomes for similar instances. This allows the use, for fairness purposes, of mechanisms designed for differential privacy. For instance [18], propose an efficient individually fair learning algorithm based on the Exponential mechanism [43], resulting in provable loss bounds. In [111], the proposed privacy-preserving approach (ensuring DP for the sensitive attributes) can be seen as a relaxation of the strict notion of individual fairness proposed by [120]. Indeed, while the former enforces a ratio on the probabilities of different outcomes when a single example's sensitive attribute is modified, the latter enforces that the sensitive attribute is never used. Fairness through unawareness is then a strict, simple but certifiable way to ensure sensitive attribute privacy.

4.2.5 | Privacy and Fairness Can Enhance Each Other in Particular Setups

Khalili et al. [121] consider the particular setting in which a pre-trained model generates qualification scores for a set of applicants. These scores are then used to determine a fixed number of candidates that will be selected by the process (e.g., for a grant, a job, etc.). They show that the Exponential mechanism can be used to perform the selection given the qualification scores, to both enforce DP for the selection process and improve fairness (here equal opportunity). Under some conditions regarding the properties of the subgroups, the proposed approach can make the selection procedure perfectly fair. Other notions of privacy can also have different interactions with fairness definitions. For instance [122], studies the context of itemset mining, in which given a dataset, the objective is to mine frequent patterns. Then, the author shows that anonymizing the data to achieve t-closeness with carefully chosen parameters implies popular group fairness notions. Finally, it is possible to perform statistically significant fairness audits using differentially private sensitive attributes, taking into account the added noise [123].

Other work [124] also considers frequent patterns discovery, and propose two-step algorithms to jointly address non-discrimination (fairness) and privacy. More precisely, they first apply a privacy-preserving mechanism, before using data sanitization methods to enforce non-discrimination. Indeed, considering either k-anonymity or DP, they theoretically prove that the privacy guarantees are not affected by the later fairness-enhancing stage. On the contrary, they observe that applying privacy-preserving mechanisms on a sanitized data could alter the resulting patterns' fairness, either increasing or decreasing discrimination depending on the considered scenario (in line with the aforementioned tensions). Importantly, they empirically note that the utility loss incurred by jointly enforcing fairness and privacy is only marginally higher than that of enforcing privacy only. This result highlights a synergy between the two desiderata, in which the former privacy-enhancing step sometimes also improves fairness, overall leading to a smaller utility drop from the later discrimination sanitizing step. This trend is valid for both k-anonymity and DP, although the later leads to a higher utility cost.

More recently [125], have conducted an empirical study on the impact of various syntactic anonymization models—namely, k-anonymity, ℓ-diversity and t-closeness—on both statistical and individual fairness. Their findings indicate that such anonymization techniques often degrade statistical fairness, with metrics like statistical parity and equalized odds showing significant deterioration. In contrast, individual fairness tends to improve under anonymization, though the extent of this improvement depends on the specific anonymization method and fairness metric considered.

5 | Interpretability and Privacy

In this section, we first discuss some tensions between interpretability and privacy. Although these notions inherently conflict, we then highlight synergies between them, before summarizing existing frameworks addressing them jointly.

5.1 | Tensions

5.1.1 | Interpretability/Explainability and Privacy Conceptually Have Antagonist Goals

While interpretability and privacy protection are both important requirements for responsible machine learning, they intrinsically pursue contrasting objectives [7]. Indeed, on one hand, interpretability aims at providing more information to enhance users' understanding of a model's behavior. On the other hand, privacy requires a tight control of the leaked information, often obfuscating part of it to protect individuals' data. Jointly addressing both desiderata hence necessitates some form of arbitration [126].

5.1.2 | Explainability Tools Can Be Used With the Purpose of Designing Attacks Against Machine Learning Models

Tools from explainable AI can be leveraged by malicious entities to perform more effective attacks against machine learning based systems. For instance [127], studied malware detection models, that are usually trained on crowd-sourced data to distinguish between malicious softwares (malwares) and legitimate ones. The authors investigated backdoor poisoning attacks, in which an

attacker injects carefully chosen datapoints to the crowd-sourced training set, resulting in its chosen malware being wrongly classified as legitimate by the detection model. In this context, they leverage Shapley values to identify highly effective features and their values, and efficiently craft the poisoned examples. Explainable AI techniques were also leveraged to fool ML-based authentication systems, which take as input a user ID along with some fingerprinting authenticating the user uniquely. An attacker can then use perturbation-based feature explanation techniques on a local surrogate model to efficiently craft a fingerprint authenticating a desired user given its ID [128]. Again, the feature importance explanations help guiding the malicious crafting process by indicating which features most influence the decision. A counterfactual explanation framework is modified by [129] to generate adversarial examples. Counterfactual explanations of a black-box model are also used to identify the features that influence the model's decision boundaries and generate examples to conduct backdoor poisoning attacks.

5.1.3 | Post Hoc Explanations Can Be Exploited to Perform or Improve Inference Attacks

Inference attacks traditionally query a model (e.g., via a prediction API) and use its outputs to achieve their goal, for instance determining an individual's membership in the training data, reconstructing part of the training dataset, extracting the model itself, or inferring an individual's missing attributes [4, 27]. Post hoc explainability techniques, by offering explanations as additional outputs, expose a new attack surface. Several works showed that such explanations, whatever form they take (e.g., example-based, feature-based, etc.), can be leveraged to enhance the different types of privacy attacks (introduced in Section 2.4):

- *Model extraction attacks*. Gradient-based (a class of feature-based) explanations of a black-box model can be exploited by an adversary to reconstruct the underlying model [130]. In the considered setup, the adversary owns an auxiliary dataset and can query the black-box model to obtain the model's gradients as explanations for given input points. The authors have designed a near-optimal algorithm, which provably extracts the entire underlying model within a bounded number of queries, in the particular case in which it is a two-layer neural network with ReLU activations. For the general case, they design an effective heuristic inspired by previous works on standard reconstruction attacks against prediction APIs. More precisely, the attacker trains a surrogate model mimicking the black-box behavior and optimizes to match its gradients thanks to the provided explanations. The results obtained demonstrates that model extraction from gradient explanations requires orders of magnitude less queries than from the sole predictions. Another approach [131] also consider gradient-based explanations, but assume no auxiliary dataset. In such case, the data used to query the black-box and train the surrogate model is outputted by a generative model, which in turn tries to generate examples so that the surrogate disagrees with the black-box. The generative model is updated leveraging the provided gradient explanations, which dramatically reduces the required number of iterations (and queries to the black-box). Furthermore [132], show that providing counterfactual (a class of example-based) explanations (CFs) can help to realize model extraction attacks with better precision and limited number of requests. More precisely, the adversary queries the black-box model with a given attack set, and trains a surrogate using the predictions of both the attack set instances and the provided CFs. The authors empirically show that the use of the provided CFs improves the attack by both increasing the built surrogate's fidelity with respect to the black-box model, and dramatically decreasing the required number of queries. A similar approach is proposed by [129], leveraging knowledge distillation techniques to train the surrogate model, which may mitigate the potential performance harm of an architecture mismatch between the actual black-box model and the reconstructed surrogate. CFs provided by Machine-Learning-as-a-Service (MLaaS) platforms are also exploited by [133], which propose an efficient querying strategy to steal the underlying classification model. Their strategy is based on the following observation: the generated CFs usually lie close to the decision boundary, while the attack set examples do not necessarily. This leads to a "decision boundary shift issue", in which the surrogate model's decision boundary is shifted compared to that of the actual black-box. To circumvent this issue, the authors propose to generate counterfactuals for the CFs themselves, and to use them all for training the surrogate.
- *Membership inference attacks*. Feature-based explanations are leveraged by [68] to perform MIAs. More precisely, they consider both backpropagation-based (i.e., gradient-based) and perturbation-based explanations. On one hand, they demonstrate that the former leak information regarding membership, and can effectively be leveraged to perform MIAs. In particular, the explanations' variance is very informative, in the sense that explanations of training examples usually exhibit a low variance, while for unseen examples, this value can be considerably higher. This is due to the fact that for training examples, the model is usually very confident, as it was optimized on them, and small perturbations are likely to not change its predictions. On the contrary, unseen samples can be closer to the decision boundary, which results in some features having a great impact on the model's predictions (hence high gradients norms), and the resulting explanation having high variance. On the other hand, they further show using two popular perturbation-based frameworks [134, 135] that the later is more resistant to membership inference. This may be explained by the fact that perturbation-based frameworks often generate perturbed examples that lie out of the data distribution [136]. The black-box model behavior on such examples is unspecified, and so querying it with them does not provide insightful information to perform inference attacks. This also suggests that the resulting explanations may qualitatively by poorer: "privacy comes at the cost of explanation quality". Counterfactual explanations are leveraged by [129] to conduct MIAs. More precisely, the black-box model is queried with an auxiliary dataset and then the model's outputs and generated counterfactual examples are used to train a shadow model. Membership of a given example is then established by comparing the difference in

prediction probabilities between the shadow model and the actual black-box to a threshold.

- *Dataset reconstruction (and membership inference) attacks.* An example-based explainability framework based on influence functions [24] and returning influential training examples that most contribute to an example's prediction is considered by [68]. Because they explicitly reveal training points, and a training point is likely to be used to explain itself, such explanations are highly vulnerable to MIAs. Indeed, this class of explanations allows for stronger attacks, such that dataset reconstruction attacks. The authors propose two algorithms that leverage the provided example-based explanations to reconstruct (part of) the model's training set. The first algorithm is based on subspace reduction and comes with a certifiable lower bound on the number of points it discovers. Empirical evaluation shows that it can be used to retrieve most of the training dataset for high dimensional data. The second one is heuristic and offers no theoretical guarantees, but works well in practice for low dimensional data. It simply consists in using previously revealed points to reveal new points. This naturally defines an influence graph structure over the training set, in which an edge between two training examples means that one is provided as an explanation for the other. The proposed algorithm can then be used to explore entire Strongly Connected Components within this graph.
- *Model inversion attacks* [137] propose model inversion attacks that aim at reconstructing a black-box model's inputs given its outputs (here, its prediction along with some *feature-based explanation*), hence harming the privacy of test instances[5] (i.e., active users of the model). In the context of image-based tasks, they focus on different types of saliency map explanations to reconstruct the target model's input images, namely gradient-based explanations [138], influence-based explanations [139] (obtained by multiplying each input feature by its associated gradient), activation-based explanations [25] and layer-wise relevance propagation [140] (i.e., attributing pixels' importance by backpropagating neurons' relevance). The proposed attack uses an attack model, trained on an independent auxiliary dataset to predict images (given as input to the target model) given predictions and explanations (outputted by the target model). As expected, the frameworks directly using the input within the explanation computation (i.e., influence-based ones) leak more information regarding the model's inputs, hence allowing better attack results. Importantly, the paper shows that even non-explainable models can be attacked, leveraging attention transfer to build an explainable surrogate whose explanations are used to conduct the attack. With a same attack objective [141], have shown that Shapley value-based explanations provided by popular Machine Learning as a Service (MLaaS) providers can be exploited to reconstruct the private model inputs. They provide an information-theoretical analysis of the relationship between an example and its associated Shapley values, and demonstrate that an adversary can always infer useful information about the former using the later. This analysis also holds for sampling-based Shapley-values, which are commonly computed as an efficient approximation of the exact Shapley values. They then studied two distinct adversarial settings, and have shown that even an adversary with no background knowledge can reconstruct most of the private model's input examples given only its outputs and explanations.
- *(Sensitive) attribute inference attacks.* Sensitive attribute inference attacks can leverage feature-based model explanations, computed either with backpropagation-based or perturbation-based methods [142]. The authors consider the two scenarios where the sensitive attribute is (or not) used for training the model and for inference. In both studied scenarios, the adversary leverages an auxiliary dataset to train an attack model to predict an example's sensitive attribute given only the outputs of the target model (prediction and explanation). They empirically show that their attack is able to leverage such explanations to perform attribute inference attack. Furthermore, they suggest that model explanations lead to higher attack success compared to model predictions, hence constituting a stronger attack surface to exploit.

5.1.4 | Interpretable Models Inherently Leak Information Regarding Their Training Data

The approach of [143] exploits the structure of a trained decision tree to reconstruct a probabilistic version of its training set. It is generalized by [144] to handle more generic types of knowledge and reconstruct probabilistic datasets from other types of interpretable models. Both works use tools from the information theory to precisely quantify the amount of knowledge interpretable models encode, through their structure, regarding their training data. By formulating a dataset reconstruction attack as a constraint programming optimization model, recent work [145] has shown that random forests can fully leak their training data through the structure of their trees. The use of bootstrap aggregating when fitting the forest helps mitigate the success of such attacks, but nonetheless the theoretical sampling probabilities can still be exploited to achieve accurate reconstructions. A recent follow-up work adapts this attack to handle random forests complying with (pure) DP guarantees [146]. Although DP is able to mitigate the success of the reconstruction attack, it comes at a significant cost in terms of model accuracy. Furthermore, in most investigated setups, the attacker is still able to retrieve training set-specific information, evidencing potential privacy leakages.

5.1.5 | Providing Useful Yet Privacy-Protective Explanations Remains an Open Challenge

As discussed in the next subsection, differentially private explainability tools have been proposed, but always imply some trade-off between the explanation quality, the privacy guarantee and the model utility. Furthermore [130], recall that DP can help guard against attacks from prediction APIs, but it is not clear if this is a viable approach for preventing reconstruction from explanations. On the same line [68], state that "the effect of DP techniques (notably the randomness they induce) on model transparency is unknown". Furthermore, the effect of DP on the explanations' robustness and user trust are still to be investigated [132].

Overall, applying explainability techniques while preserving formal privacy guarantees is challenging. In the next subsection, we nevertheless how this could be achieved, but this implies some cost on either one aspect or the other.

5.2 | Synergies

5.2.1 | Interpretability Eases Model Audit and Can Be Leveraged for Privacy Purposes

Interpretability can be used to confirm other desiderata of ML systems, such as privacy [88]. It also makes it easier to detect possible privacy issues when building interpretable models [87]. Furthermore, this auditable nature is particularly appreciated in the area of ML-based cybersecurity systems [147]. Indeed, machine learning models have shown great abilities to detect abnormal behaviors or intrusions. However, their black-box nature and lack of certification can be problematic as it possibly introduces weaknesses inside the security system. By providing an understanding of the underlying mechanisms and reasoning of the model, interpretability techniques can be helpful to detect overfitting, or in cases in which the model captures noise or inaccurate values in the data. This allows deploying more trustworthy models, but also helps the administrators identify potential breaches.

5.2.2 | Interpretability Can Be Conciliated With Privacy With Some Trade-Offs

Friedman and Schuster [148] study data mining with DP guarantees, considering decision tree learning as an illustrative task. They demonstrate that the design of the privacy preserving mechanism is crucial, and that there is a huge difference in terms of model utility and required sample size between a naive implementation using a general purpose privacy preserving data interface and a task-specific differentially private learning algorithm. Their empirical study demonstrates the ability of their proposed algorithm to learn differentially private decision trees with reasonable cost in terms of accuracy. Several other works also tackled differentially private decision tree building, as summarized by [149]. Locally Linear Maps (LLMs) are studied by [150] and consist in a linear combination of logistic regressions for each possible class. Such interpretable models are suitable to provide local explanations (using the appropriate LLM) but also global ones, as the coefficients of each class's LLMs provide insights regarding which features really matter to it. The authors propose a procedure to learn LLMs under DP, leveraging mechanisms from the `DP-SGD` framework [46]. They empirically observe a trade-off between the privacy guarantee and the model's accuracy and interpretability.

5.2.3 | Post Hoc Explainability Can Be Conciliated With Privacy With Some Trade-Offs

Quantitative Input Influence (QII) is a framework leveraging Shapley values to provide feature-based explanations quantifying the influence of input features over the model's predictions [106]. As such measures may leak information regarding individual users, the authors introduce a mechanism to generate differentially private explanations to the so-called transparency queries. Providing pure DP guarantees, it consists in adding Laplace noise to the query answers, scaled to the query function sensitivity. As the proposed measures generally have low sensitivity, the amount of added noise remains reasonable which results in relatively small average utility losses. Nonetheless, for some types of explanations with exceptionally high sensitivity, the amount of noise added may significantly harm their utility. A method to generate differentially private feature-based explanations (i.e., local linear surrogates) of a black-box model is introduced by [107]. In their framework, the explanations are computed using a differentially private gradient descent leveraging the Gaussian mechanism. They further proposed an adaptive mechanism, reducing the spending of the privacy budget by leveraging the explanations to previous queries when computing a new one. Using tabular, text and image data, they empirically observe that the explanations' quality degrade while the privacy guarantees tighten [151] investigated the impact of a model's differential privacy on the quality of post hoc explanations (saliency maps [25]) of this model and on its utility, considering either local DP (classical learning algorithm applied on DP data) or global DP (differentially private training algorithm). In both cases, the explanations are also differentially private due to the post-processing property (cf. Section 2.4). Handling either general or medical imaging applications, they have learnt neural networks under different DP budgets and evaluate the quality of post hoc explanations of their predictions using two metrics from the literature. In a nutshell, these metrics aim at quantifying how much the regions highlighted by explanation maps actually account for the explained decisions. The experimental results show that these metrics degrade while the privacy budget is tightened. Furthermore, they suggest the existence of a three dimensional trade-off space between privacy, explanation quality and model accuracy.

To face the explanation-guided backdoor poisoning attack studied by [127] (and discussed in Section 5.1) [152], proposed to generate Locally Differentially Private explanations. By randomly perturbating the top-k features in the generated feature-based explanations, the mechanism is shown to mitigate the success of the attack. An approach to generate robust counterfactual explanations for differentially private Support Vector Machines (SVMs) is designed by [153]. More precisely, privacy is achieved by adding Laplace noise to the SVMs' weights, and classical counterfactual explanation frameworks may generate counterfactuals that allow to cross the classifier's noisy boundaries, but not to actually change the example's class in real-life. To address this issue, they instead generate robust counterfactual explanations by solving an optimization problem with probabilistic constraints. In practice, the generated counterfactuals require more and more changes to the example as the privacy level tightens, to ensure that its classification changes with respect to the (unknown) non-private classifier. Again, this illustrates the trade-off between explanations quality and privacy protection. In the context of federated learning [154], have also noticed that DP can alter the meaningfulness of gradient-based explanations. They propose an adaptive mechanism still providing DP guarantees but injecting noise within the model's parameters in a manner aimed at preserving the quality of gradient-based explanations. Finally, recent work also studied DP for counterfactual explanations [155]. The approach consists in using

an autoencoder trained in a differentially private manner to build noisy class prototypes, which can then be leveraged to generate the counterfactuals. DP has also been applied to protect training data in the context of algorithmic recourse. Specifically [156], first perform a DP clustering of the training data, from which they construct a graph whose nodes correspond to cluster centers and whose edges encode actionability constraints. This graph is then used at inference time to generate counterfactual explanations and step-by-step recourse paths, with no additional privacy loss thanks to the post-processing property of differential privacy. Finally, beyond DP, other privacy-preserving strategies—including, but not limited to, syntactic models of anonymity—have also been explored to protect post hoc explanations. For a comprehensive review of such approaches, we refer the reader to the recent survey by [157].

6 | Conclusion

We have seen throughout this paper that while fairness, interpretability and privacy are three important dimensions of responsible ML, they often conflict in different ways, both theoretically and empirically. A thorough characterization of these tensions is crucial to support informed design choices by stakeholders in practice. Among other challenges, precisely accounting for the (direct or indirect) privacy cost introduced by fairness mechanisms, as well as systematically and rigorously quantifying the privacy risk posed by post hoc explanations, remain important and largely open research directions. Nonetheless, we have also identified synergies, which suggests that a careful design can sometimes lead to improving them jointly with a reduced impact on utility. However, this considerably increases the complexity of the learning process while requiring an in-depth analysis of the used techniques. Throughout this paper, we have highlighted several interesting works taking advantage of these synergies to conciliate two of our three pillars. These insightful examples include modifying or analyzing the distribution of the noise added by privacy-preserving techniques to improve [108] or certify [117] statistical fairness, leveraging fairness constraints to enhance the learning of interpretable models through effective pruning mechanisms [58] or leveraging explainability tools to detect privacy leakages [147]. Previous work has also demonstrated the potential of leveraging differential privacy tools to promote individual fairness, highlighting a promising avenue for future research. In particular, extensions of DP—such as metric DP [158], which generalizes the notion of neighboring databases to metric spaces—offer powerful tools for enforcing individual fairness.

Nevertheless, compromises usually have to be made. Generally speaking, learning a model with non-trivial utility and satisfying our three desiderata requires a thorough theoretical formulation, being aware of the existing tensions as well as of common techniques to mitigate them. We graphically summarize them in Figures 1, 2, and 3. More precisely, for each pair of desiderata, we position the identified compatibilities, synergies and tensions within a two-dimensional space based on whether they tend to enhance or harm each desiderata. We believe that such a summary of these interplays can be beneficial for stakeholders to be aware of the possible tensions they may have to face, and of the existing compatibilities and synergies they can leverage to develop trustworthy yet accurate machine learning models. This is crucial since a naive joint implementation of these desiderata would likely result in suboptimal trade-offs, and some scenarios may even lead to mutually worsening them. We also aim at encouraging research regarding these interplays—and to summarize them in a systematic manner so that they benefit the field.

Finally, it is crucial to promote an interdisciplinary approach, for computer scientists to ensure that the metrics they optimize for actually match legal and ethical requirements. This is a particularly challenging aspect: ethical analysis is often strongly context-dependent while genericity is a common objective in computer science. In addition, not all legal and ethical notions can easily be implemented and quantified using mathematical formulas. It is hence necessary to verify the alignment of the notions we use with the concepts we target, for the development of ML systems that can be trusted and that do not harm the society. There exist several works specifically considering these aspects, such as that of [159] which reviews critics of popular fairness-enhancing approaches from an interdisciplinary perspective.

Acknowledgments

This work was partially supported by the Canada Research Chairs Program (Privacy-preserving and ethical analysis of Big Data Canada Research Chair) and by the LabEx CIMI (ANR-11-LABX-0040).

Data Availability Statement

The authors have nothing to report.

Endnotes

[1] https://gdpr-info.eu/.

[2] https://artificialintelligenceact.eu/.

[3] https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/differential-privacy.html.

[4] That is, pure $(\epsilon,0)$-DP or approximate (ϵ,δ)-DP with δ cryptographically small, which are the meaningful and typical configurations of DP.

[5] This differs from the previously mentioned reconstruction attacks. Indeed, in reconstruction attacks, the objective of the adversary is to infer information regarding the model's training data. In the discussed model inversion attacks, the objective is to gain information about the examples provided to the model at inference time, by only observing the model's outputs (cf., Section 2.4).

References

1. P. Voigt and A. Von dem Bussche, *The Eu General Data Protection Regulation (Gdpr). A Practical Guide*, vol. 10(3152676), 1st ed. (Springer International Publishing, 2017), 10–5555.

2. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys* 54, no. 6 (2022): 115:1–115:35.

3. L. Deck, J. L. Müller, C. Braun, D. Zipperling, and N. Kühl, "Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness," (2024) arXiv preprint arXiv:240320089.

4. E. De Cristofaro, "An Overview of Privacy in Machine Learning," (2020), arXiv preprint arXiv:200508679.

5. S. Barocas, M. Hardt, and A. Narayanan, "Fairness and Machine Learning," (2019), fairmlbook.org.

6. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys* 51, no. 5 (2018): 1–42.

7. T. Datta, D. Nissani, M. Cembalest, A. Khanna, H. Massa, and J. P. Dickerson, "Position: Tensions Between the Proxies of Human Values in AI," in *First IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2023* (IEEE, 2023).

8. S. L. Garfinkel, J. M. Abowd, and S. Powazek, "Issues Encountered Deploying Differential Privacy," in *Proceedings of the 2018 Workshop on Privacy in the Electronic Society* (Association for Computing Machinery, 2018), 133–137.

9. S. Wachter, B. Mittelstadt, and C. Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI," *Computer Law & Security Review* 41 (2021): 105567.

10. K. Meding, "It's Complicated. The Relationship of Algorithmic Fairness and Non-Discrimination Regulations in the EU AI Act," (2025), arXiv preprint arXiv:250112962.

11. D. Pujol, R. McKenna, S. Kuppam, M. Hay, A. Machanavajjhala, and G. Miklau, "Fair Decision Making Using Privacy-Protected Data," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* 2020* (ACM, 2020), 189–199.

12. F. Fioretto, C. Tran, P. V. Hentenryck, and K. Zhu, "Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022* (International Joint Conferences on Artificial Intelligence Organization, 2022), 5470–5477.

13. T. Telford, "Apple Card Algorithm Sparks Gender Bias Allegations Against Goldman Sachs," *Washington Post* (2019): 11, https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/.

14. W. Knight, "The Apple Card Didn't'see'gender—And That's the Problem," *Wired* (2019): 19, https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/.

15. T. H. Rafi, F. A. Noor, T. Hussain, and D. Chae, "Fairness and Privacy Preserving in Federated Learning: A Survey," *Information Fusion* 105 (2024): 102198.

16. J. Schöffer, "On the Interplay of Transparency and Fairness in AI-Informed Decision-Making," in *PhD Thesis, Karlsruher Institut für Technologie (KIT)* (Karlsruher Institut für Technologie (KIT), 2023).

17. S. Verma and J. Rubin, "Fairness Definitions Explained," in *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018* (ACM, 2018), 1–7.

18. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness Through Awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS 2012* (ACM, 2012), 214–226.

19. N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding Discrimination Through Causal Reasoning," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NeurIPS 2017* (Curran Associates Inc, 2017), 656–666.

20. R. K. E. Bellamy, K. Dey, M. Hind, et al., "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias," *IBM Journal of Research and Development* 63, no. 4/5 (2019): 4:1–4:15.

21. S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning," in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, FAT* 2019* (ACM, 2019), 329–338.

22. S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," *ACM Computing Surveys* 56, no. 7 (2024): 166:1–166:38.

23. N. Burkart and M. F. Huber, "A Survey on the Explainability of Supervised Machine Learning," *Journal of Artificial Intelligence Research* 70 (2021): 245–317.

24. P. W. Koh and P. Liang, "Understanding Black-Box Predictions via Influence Functions," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017* (PMLR, 2017), 1885–1894.

25. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-Cam: Visual Explanations From Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2017), 618–626.

26. Z. C. Lipton, "The Mythos of Model Interpretability," *Queue* 16, no. 3 (2018): 31–57.

27. C. Dwork, A. Smith, T. Steinke, and J. Ullman, "Exposed! A Survey of Attacks on Private Data," *Annual Review of Statistics and Its Application* 4, no. 1 (2017): 61–84.

28. M. Rigaki and S. García, "A Survey of Privacy Attacks in Machine Learning," *ACM Computing Surveys* 56, no. 4 (2024): 101:1–101:34.

29. R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *2017 IEEE Symposium on Security and Privacy, SP 2017* (IEEE Computer Society, 2017), 3–18.

30. F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing Machine Learning Models via Prediction APIs," in *25th USENIX Security Symposium* (USENIX Security 2016 USENIX Association, 2016), 601–618.

31. M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2015), 1322–1333.

32. C. Clifton and T. Tassa, "On Syntactic Anonymity and Differential Privacy," *Transactions on Data Privacy* 6, no. 2 (2013): 161–183.

33. L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 5 (2002): 557–570.

34. P. Samarati, "Protecting Respondents' Identities in Microdata Release," *IEEE Transactions on Knowledge and Data Engineering* 13, no. 6 (2001): 1010–1027.

35. N. Li, T. Li, and S. Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007* (IEEE Computer Society, 2007), 106–115.

36. C. Dwork, F. McSherry, K. Nissim, and A. D. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006* vol. 3876 of Lecture Notes in Computer Science (Springer, 2006), 265–284.

37. J. M. Abowd, "The U.S. Census Bureau Adopts Differential Privacy," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018* (ACM, 2018), 2867.

38. A. Aktay, S. Bavadekar, G. Cossoul, et al., "Google COVID-19 Community Mobility Reports: Anonymization Process Description (Version 1.0)," (2020), arXiv preprint arXiv:200404145.

39. A. Herdagdelen, A. Dow, B. State, P. Mohassel, and A. Pompe, "Protecting Privacy in Facebook Mobility Data During the COVID-19 Response," (2020).

40. Team ADP, "Learning With Privacy at Scale," *Apple Machine Learning Journal* (2017), https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

41. C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science* 9, no. 3–4 (2014): 211–407.

42. J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett, "Functional Mechanism: Regression Analysis Under Differential Privacy," *Proceedings of the VLDB Endowment* 5, no. 11 (2012): 1364–1375.

43. F. McSherry and K. Talwar, "Mechanism Design via Differential Privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007)* (IEEE Computer Society, 2007), 94–103.

44. Z. Ji, Z. C. Lipton, and C. Elkan, "Differential Privacy and Machine Learning: A Survey and Review," (2014), arXiv preprint arXiv:14127584.

45. M. Gong, Y. Xie, K. Pan, K. Feng, and A. K. Qin, "A Survey on Differentially Private Machine Learning [Review Article]," *IEEE Computational Intelligence Magazine* 15, no. 2 (2020): 49–64.

46. M. Abadi, A. Chu, I. J. Goodfellow, et al., "Deep Learning With Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2016), 308–318.

47. N. Papernot, M. Abadi, Ú. Erlingsson, I. J. Goodfellow, and K. Talwar, "Semi-Supervised Knowledge Transfer for Deep Learning From Private Training Data," in *5th International Conference on Learning Representations, ICLR 2017* (2017).

48. N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable Private Learning With PATE," in *6th International Conference on Learning Representations, ICLR 2018* (2018).

49. N. Ponomareva, H. Hazimeh, A. Kurakin, et al., "How to DP-Fy ML: A Practical Guide to Machine Learning With Differential Privacy," *Journal of Artificial Intelligence Research* 77 (2023): 1113–1201.

50. S. Agarwal, "Trade-Offs Between Fairness and Interpretability in Machine Learning," in *IJCAI 2021 Workshop on AI for Social Good* (2021).

51. J. Kleinberg and S. Mullainathan, "Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability," in *Proceedings of the 2019 ACM Conference on Economics and Computation* (Association for Computing Machinery, 2019), 807–808.

52. G. K. Dziugaite, S. Ben-David, and D. M. Roy, "Enforcing Interpretability and Its Statistical Impacts: Trade-Offs Between Accuracy and Interpretability," (2020), arXiv preprint arXiv:201013764.

53. G. Dai, P. Ravishankar, R. Yuan, D. B. Neill, and E. Black, "Be Intentional About Fairness!: Fairness, Size, and Multiplicity in the Rashomon Set," (2025), arXiv preprint arXiv:250115634.

54. L. Langlade, J. Ferry, G. Laberge, and T. Vidal, "Fairness and Sparsity Within Rashomon Sets: Enumeration-Free Exploration and Characterization," (2025), arXiv preprint arXiv:250205286.

55. S. Jabbari, H. C. Ou, H. Lakkaraju, and M. Tambe, "An Empirical Study of the Trade-Offs Between Interpretability and Fairness," in *ICML 2020 Workshop on Human Interpretability in Machine Learning* (2020).

56. N. Jo, S. Aghaei, J. Benson, A. Gómez, and P. Vayanos, "Learning Optimal Fair Decision Trees: Trade-Offs Between Interpretability, Fairness, and Accuracy," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023* (ACM, 2023), 181–192.

57. C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges," *Statistics Surveys* 16 (2022): 1–85.

58. U. Aïvodji, J. Ferry, S. Gambs, M. Huguet, and M. Siala, "Leveraging Integer Linear Programming to Learn Optimal Fair Rule Lists," in *Integration of Constraint Programming, Artificial Intelligence, and Operations Research - 19th International Conference, CPAIOR 2022* vol. 13292 of Lecture Notes in Computer Science (Springer, 2022), 103–119.

59. J. Dai, S. Upadhyay, S. H. Bach, and H. Lakkaraju, "What Will It Take to Generate Fairness-Preserving Explanations?," (2021), arXiv preprint arXiv:210613346.

60. U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp, "Fairwashing: The Risk of Rationalization," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Vol. 97 of Proceedings of Machine Learning Research* (PMLR, 2019), 161–170.

61. M. M. Manerba and R. Guidotti, "Investigating Debiasing Effects on Classification and Explainability," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES 2021* (ACM, 2022), 468–478.

62. J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, "Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment," in *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Association for Computing Machinery, 2019), 275–285.

63. C. Wang, B. Han, B. Patel, and C. Rudin, "In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction," *Journal of Quantitative Criminology* 39, no. 2 (2023): 519–581.

64. F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification Without Discrimination," *Knowledge and Information Systems* 33, no. 1 (2012): 1–33.

65. R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013* (PMLR, 2013), 325–333.

66. G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NeurIPS 2017* (Curran Associates Inc., 2017).

67. R. Shokri, M. Strobel, and Y. Zick, "Exploiting Transparency Measures for Membership Inference: A Cautionary Tale," in *The AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI)*, vol. 13 (2020).

68. R. Shokri, M. Strobel, and Y. Zick, "On the Privacy Risks of Model Explanations," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery, 2021), 231–241.

69. J. Dai, S. Upadhyay, U. Aïvodji, S. H. Bach, and H. Lakkaraju, "Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post Hoc Explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES 2021* (ACM, 2022), 203–214.

70. A. Balagopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi, "The Road to Explainability Is Paved With Bias: Measuring the Fairness of Explanations," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2022* (ACM, 2022), 1194–1206.

71. B. Ustun, A. Spangher, and Y. Liu, "Actionable Recourse in Linear Classification," in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, FAT* 2019* (ACM, 2019), 10–19.

72. S. Sharma, J. Henderson, and J. Ghosh, "CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES 2020* (ACM, 2020), 166–172.

73. V. Gupta, P. Nokhiz, C. D. Roy, and S. Venkatasubramanian, "Equalizing Recourse Across Groups," (2019), arXiv preprint arXiv:190903166.

74. A. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations," *ACM Computing Surveys* 55, no. 5 (2023): 95:1–95:29.

75. H. Lakkaraju and O. Bastani, "'How Do I Fool You?': Manipulating User Trust via Misleading Black Box Explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES 2020. AIES 2020* (ACM, 2020), 79–85.

76. H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Faithful and Customizable Explanations of Black Box Models," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019* (ACM, 2019), 131–138.

77. U. Aïvodji, H. Arai, S. Gambs, and S. Hara, "Characterizing the Risk of Fairwashing," in *Advances in Neural Information Processing Systems*

34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021 (Curran Associates Inc., 2021), 14822–14834.

78. D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES 2020* (ACM, 2020), 180–186.

79. D. Slack, S. Hilgard, S. Singh, and H. Lakkaraju, "Feature Attributions and Counterfactual Explanations Can be Manipulated," (2021), arXiv preprint arXiv:210612563.

80. J. Heo, S. Joo, and T. Moon, "Fooling Neural Network Interpretations via Adversarial Model Manipulation," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019* (Curran Associates Inc., 2019), 2921–2932.

81. B. Dimanov, U. Bhatt, M. Jamnik, and A. Weller, "You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods," in *24th European Conference on Artificial Intelligence, ECAI 2020, Vol. 325 of Frontiers in Artificial Intelligence and Applications* (IOS Press, 2020), 2473–2480.

82. D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton, "Learning to Deceive With Attention-Based Explanations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020 Association for Computational Linguistics* (Association for Computational Linguistics, 2020), 4782–4793.

83. D. Slack, A. Hilgard, H. Lakkaraju, and S. Singh, "Counterfactual Explanations Can be Manipulated," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021* (Curran Associates Inc., 2021), 62–75.

84. X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang, "Interpretable Deep Learning Under Fire," in *29th USENIX Security Symposium, USENIX Security 2020* (USENIX Association, 2020), 1659–1676.

85. G. Laberge, U. Aïvodji, S. Hara, M. Marchand, and F. Khomh, "Fooling SHAP With Stealthily Biased Sampling," in *11th International Conference on Learning Representations, ICLR 2023* (2023).

86. E. L. Merrer and G. Trédan, "The Bouncer Problem: Challenges to Remote Explainability," (2019), arXiv preprint arXiv:191001432.

87. C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* 1, no. 5 (2019): 206–215.

88. F. Doshi-Velez and B. Kim, "A Roadmap for a Rigorous Science of Interpretability," (2017), arXiv preprint arXiv:170208608.

89. T. Begley, T. Schwedes, C. Frye, and I. Feige, "Explainability for Fair Machine Learning," (2020), arXiv preprint arXiv:201007389.

90. N. Kilbertus, A. Gascón, M. J. Kusner, M. Veale, K. P. Gummadi, and A. Weller, "Blind Justice: Fairness With Encrypted Sensitive Attributes," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Vol. 80 of Proceedings of Machine Learning Research* (PMLR, 2018), 2635–2644.

91. 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, UMAP 2019* (ACM, 2019), 309–315.

92. S. Agarwal, "Trade-Offs Between Fairness and Privacy in Machine Learning," in *IJCAI 2021 Workshop on AI for Social Good* (2021).

93. B. Kulynych, M. Yaghini, G. Cherubin, M. Veale, and C. Troncoso, "Disparate Vulnerability to Membership Inference Attacks," in *Privacy-Enhancing Technologies Symposium, PETS 2022* (2022), 460–480.

94. M. D. Ekstrand, R. Joshaghani, and H. Mehrpouyan, "Privacy for All: Ensuring Fair and Equitable Privacy Protections," in *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, FAT* 2018* (PMLR, 2018), 35–47.

95. H. Chang and R. Shokri, "On the Privacy Risks of Algorithmic Fairness," in *IEEE European Symposium on Security and Privacy, EuroS&P 202* (IEEE, 2021), 292–303.

96. H. Hu and C. Lan, "Inference Attack and Defense on the Distributed Private Fair Learning Framework," in *The AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI)* (2020).

97. J. Ferry, U. Aïvodji, S. Gambs, M. J. Huguet, and M. Siala, "Exploiting Fairness to Enhance Sensitive Attributes Reconstruction," in *First IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2023* (IEEE, 2023).

98. E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential Privacy Has Disparate Impact on Model Accuracy," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019* (Curran Associates Inc., 2019), 15453–15462.

99. A. Uniyal, R. Naidu, S. Kotti, et al., "DP-SGD vs PATE: Which Has Less Disparate Impact on Model Accuracy?," (2021), arXiv preprint arXiv:210612576.

100. T. Farrand, F. Mireshghallah, S. Singh, and A. Trask, "Neither Private nor Fair: Impact of Data Imbalance on Utility and Fairness in Differential Privacy," in *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice, PPMLP 2020* (ACM, 2020), 15–19.

101. V. M. Suriyakumar, N. Papernot, A. Goldenberg, and M. Ghassemi, "Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2021* (ACM, 2021), 723–734.

102. C. Tran, M. H. Dinh, K. Beiter, and F. Fioretto, "A Fairness Analysis on Private Aggregation of Teacher Ensembles," (2021), arXiv preprint arXiv:210908630.

103. D. Xu, W. Du, and X. Wu, "Removing Disparate Impact on Model Accuracy in Differentially Private Stochastic Gradient Descent," in *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2021* (ACM, 2021), 1924–1932.

104. T. Zhang, T. Zhu, K. Gao, W. Zhou, and P. S. Yu, "Balancing Learning Model Privacy, Fairness, and Accuracy With Early Stopping Criteria," *IEEE Transactions on Neural Networks and Learning Systems* 34, no. 9 (2023): 5557–5569.

105. K. Koch and M. Soll, "No Matter How You Slice It: Machine Unlearning With SISA Comes at the Expense of Minority Classes," in *First IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2023* (IEEE, 2023).

106. A. Datta, S. Sen, and Y. Zick, "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments With Learning Systems," in *IEEE Symposium on Security and Privacy, SP 2016* (IEEE Computer Society, 2016), 598–617.

107. N. Patel, R. Shokri, and Y. Zick, "Model Explanations With Differential Privacy," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2022* (ACM, 2022), 1895–1904.

108. D. Xu, S. Yuan, and X. Wu, "Achieving Differential Privacy and Fairness in Logistic Regression," in *Companion of the 2019 World Wide Web Conference, WWW 2019* (ACM, 2019), 594–599.

109. J. Ding, X. Zhang, X. Li, J. Wang, R. Yu, and M. Pan, "Differentially Private and Fair Classification via Calibrated Functional Mechanism," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020* (AAAI Press, 2020), 622–629.

110. C. Tran, F. Fioretto, and P. V. Hentenryck, "Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach," in *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021* (AAAI Press, 2021), 9932–9939.

111. M. Jagielski, M. J. Kearns, J. Mao, et al., "Differentially Private Fair Learning," in *Proceedings of the 36th International Conference on Machine*

Learning, ICML 2019, Vol. 97 of Proceedings of Machine Learning Research (PMLR, 2019), 3000–3008.

112. M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, NeurIPS 2016*, vol. 29 (Curran Associates, Inc, 2016).

113. A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. M. Wallach, "A Reductions Approach to Fair Classification," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Vol. 80 of Proceedings of Machine Learning Research* (PMLR, 2018), 60–69.

114. H. Mozannar, M. Ohannessian, and N. Srebro, "Fair Learning With Private Demographic Data," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020* (PMLR, 2020), 7066–7075.

115. H. H. Arcolezi, K. Makhlouf, and C. Palamidessi, "(Local) Differential Privacy Has NO Disparate Impact on Fairness," in *Proceedings of the 37th Annual IFIP WG 11.3 Conference, DBSec 2023* vol. 13942 of Lecture Notes in Computer Science (Springer, 2023), 3–21.

116. A. S. de Oliveira, C. Kaplan, K. Mallat, and T. Chakraborty, "An Empirical Analysis of Fairness Notions Under Differential Privacy," (2023), arXiv preprint arXiv:230202910.

117. K. Tran, F. Fioretto, I. Khalil, M. T. Thai, L. T. X. Phan, and N. Phan, "FairDP: Achieving Fairness Certification With Differential Privacy," in *Third IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2025* (IEEE, 2025), 956–976.

118. P. Mangold, M. Perrot, A. Bellet, and M. Tommasi, "Differential Privacy Has Bounded Impact on Fairness in Classification," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Vol. 202 of Proceedings of Machine Learning Research* (PMLR, 2023), 23681–23705.

119. B. Imana, A. Korolova, and J. S. Heidemann, "Having Your Privacy Cake and Eating It Too: Platform-Supported Auditing of Social Media Algorithms for Public Interest," *Proceedings of the ACM on Human-Computer Interaction* 7, no. CSCW1 (2023): 1–33.

120. A. Ignatiev, M. C. Cooper, M. Siala, E. Hebrard, and J. Marques-Silva, "Towards Formal Fairness in Machine Learning," in *International Conference on Principles and Practice of Constraint Programming, CP 2020* (Springer, 2020), 846–867.

121. M. M. Khalili, X. Zhang, M. Abroshan, and S. Sojoudi, "Improving Fairness and Privacy in Selection Problems," in *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021* (AAAI Press, 2021), 8092–8100.

122. S. Ruggieri, "Data Anonymity Meets Non-Discrimination," in *2013 IEEE 13th International Conference on Data Mining Workshops* (IEEE, 2013), 875–882.

123. R. Friedberg and R. Rogers, "Privacy Aware Experimentation Over Sensitive Groups: A General Chi Square Approach," in *Algorithmic Fairness Through the Lens of Causality and Privacy Workshop, AFCP 2022, Vol. 214 of Proceedings of Machine Learning Research* (PMLR, 2022), 23–66.

124. S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti, "Discrimination- and Privacy-Aware Patterns," *Data Mining and Knowledge Discovery* 29, no. 6 (2015): 1733–1782.

125. H. H. Arcolezi, M. Alishahi, A. A. Bendoukha, and N. Kaaniche, "Fair Play for Individuals, Foul Play for Groups? Auditing Anonymization's Impact on ML Fairness," (2025), arXiv preprint arXiv:250507985.

126. D. Banisar, *The Right to Information and Privacy: Balancing Rights and Managing Conflicts* (World Bank, 2011).

127. G. Severi, J. Meyer, S. E. Coull, and A. Oprea, "Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers," in *30th USENIX Security Symposium, USENIX Security 2021 USENIX Association* (USENIX Association, 2021), 1487–1504.

128. W. Garcia, J. I. Choi, S. K. Adari, S. Jha, and K. R. B. Butler, "Explainable Black-Box Attacks Against Model-Based Authentication," (2018), arXiv preprint arXiv:181000024.

129. A. Kuppa and N. Le-Khac, "Adversarial XAI Methods in Cybersecurity," *IEEE Transactions on Information Forensics and Security* 16 (2021): 4924–4938.

130. S. Milli, L. Schmidt, A. D. Dragan, and M. Hardt, "Model Reconstruction From Model Explanations," in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, FAT* 2019* (ACM, 2019), 1–9.

131. T. Miura, S. Hasegawa, and T. Shibahara, "MEGEX: Data-Free Model Extraction Attack Against Gradient-Based Explainable AI," in *Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep Learning Systems* (Association for Computing Machinery, 2024), 56–66.

132. U. Aïvodji, A. Bolot, and S. Gambs, "Model Extraction From Counterfactual Explanations," (2020), arXiv preprint arXiv:200901884.

133. Y. Wang, H. Qian, and C. Miao, "DualCF: Efficient Model Extraction Attack From Counterfactual Explanations," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2022* (ACM, 2022), 1318–1329.

134. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should i Trust You?" Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2016), 1135–1144.

135. D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "SmoothGrad: Removing Noise by Adding Noise," (2017), arXiv preprint arXiv:170603825.

136. I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems With Shapley-Value-Based Explanations as Feature Importance Measures," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020* (PMLR, 2020), 5491–5500.

137. X. Zhao, W. Zhang, X. Xiao, and B. Lim, "Exploiting Explanations for Model Inversion Attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2021), 682–692.

138. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," in *2nd International Conference on Learning Representations, ICLR 2014* (2014).

139. H. G. Ramaswamy, "Ablation-Cam: Visual Explanations for Deep Convolutional Network via Gradient-Free Localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (IEEE, 2020), 983–991.

140. S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS One* 10, no. 7 (2015): e0130140.

141. X. Luo, Y. Jiang, and X. Xiao, "Feature Inference Attack on Shapley Values," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022* (ACM, 2022), 2233–2247.

142. V. Duddu and A. Boutet, "Inferring Sensitive Attributes From Model Explanations," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM 2022* (ACM, 2022), 416–425.

143. S. Gambs, A. Gmati, and M. Hurfin, "Reconstruction Attack Through Classifier Analysis," in *Data and Applications Security and Privacy XXVI-26th Annual IFIP WG 11.3 Conference, DBSec 2012* vol. 7371 of Lecture Notes in Computer Science (Springer, 2012), 274–281.

144. J. Ferry, U. Aïvodji, S. Gambs, M. J. Huguet, and M. Siala, "Probabilistic Dataset Reconstruction From Interpretable Models," in *2nd IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2024* (IEEE, 2024).

145. J. Ferry, R. Fukasawa, T. Pascal, and T. Vidal, "Trained Random Forests Completely Reveal Your Dataset," in *Proceedings of the 41st International Conference on Machine Learning, ICML 2024, 235 of Proceedings of Machine Learning Research* (PMLR, 2024), 13545–13569.

146. A. Gorgé, J. Ferry, S. Gambs, and T. Vidal, "Training Set Reconstruction From Differentially Private Forests: How Effective Is DP?," (2025), arXiv preprint arXiv:250205307.

147. G. Srivastava, R. H. Jhaveri, S. Bhattacharya, et al., "XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions," (2022), arXiv preprint arXiv:220603585.

148. A. Friedman and A. Schuster, "Data Mining With Differential Privacy," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2010), 493–502.

149. S. Fletcher and M. Z. Islam, "Decision Tree Classification With Differential Privacy: A Survey," *ACM Computing Surveys* 52, no. 4 (2019): 83:1–83:33.

150. F. Harder, M. Bauer, and M. Park, "Interpretable and Differentially Private Predictions," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020* (AAAI Press, 2020), 4083–4090.

151. R. Naidu, A. Priyanshu, A. Kumar, S. Kotti, H. Wang, and F. Mireshghallah, "When Differential Privacy Meets Interpretability: A Case Study," (2021), arXiv preprint arXiv:210613203.

152. T. D. T. Nguyen, P. Lai, H. Phan, and M. T. Thai, "XRand: Differentially Private Defense Against Explanation-Guided Attacks," *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (2023): 11873–11881.

153. R. Mochaourab, S. Sinha, S. Greenstein, and P. Papapetrou, "Robust Explanations for Private Support Vector Machines," in *International Conference on Machine Learning (ICML 2021), Workshop on Socially Responsible Machine Learning* (2021).

154. Z. Li, H. Chen, Z. Ni, and H. Shao, "Balancing Privacy Protection and Interpretability in Federated Learning," (2023), arXiv preprint arXiv:230208044.

155. F. Yang, Q. Feng, K. Zhou, J. Chen, and X. Hu, "Differentially Private Counterfactuals via Functional Mechanism," (2022), arXiv preprint arXiv: 220802878.

156. S. Pentyala, S. Sharma, S. Kariyappa, F. Lecue, and D. Magazzeni, "Privacy-Preserving Algorithmic Recourse," (2023), arXiv preprint arXiv:231114137.

157. T. T. Nguyen, T. T. Huynh, Z. Ren, et al., "Privacy-Preserving Explainable AI: A Survey," *Science China Information Sciences* 68, no. 1 (2025): 111101.

158. J. Imola, S. P. Kasiviswanathan, S. White, A. Aggarwal, and N. Teissier, "Balancing Utility and Scalability in Metric Differential Privacy," in *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, Vol. 180 of Proceedings of Machine Learning Research* (PMLR, 2022), 885–894.

159. L. Weinberg, "Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches," *Journal of Artificial Intelligence Research* 74 (2022): 75–109.