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

Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, Mohamed Siala

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

Julien Ferry JFERRY@LAAS.FR
Marie-José Huguet HUGUET@LAAS.FR
Mohamed Siala MSIALA@LAAS.FR
LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France

Ulrich Aïvodji ULRICH.AIVODJI@ETSMTL.CA
École de Technologie Supérieure, Montréal, Canada

Sébastien Gambs GAMBS.SEBASTIEN@UQAM.CA
Université du Québec à Montréal, Montréal, Canada

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 several legal texts, such as the European Union General Data Protection Regulation[1] (Voigt & Von dem Bussche, 2017) or the forthcoming AI Act[2].

1. https://gdpr-info.eu/
2. https://artificialintelligenceact.eu/

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 towards 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. Second, it was shown that data-driven decision-making processes can create or reproduce biases that systematically disadvantage specific individuals or groups (Mehrabi et al., 2022). Quantifying but also reducing/eliminating these biases to promote *fairness* is hence an important challenge. 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.

These three topics, namely privacy, fairness and explainability, have been extensively studied during the last decade (Cristofaro, 2020; Barocas et al., 2019; Guidotti et al., 2018) 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 (Datta et al., 2023), 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, and consider mainly classification tasks.

Positioning with respect to other surveys Other recent works survey the literature on the interactions between several of our three identified desiderata. Among others, Datta et al. (2023) 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, Fioretto et al. (2022) 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. Finally, a recent thesis (Schöffer, 2023) focuses on the interactions between transparency and fairness. It provides a deepening of (part of) our dedicated Section 3.

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. We then conclude with the identified key challenges in Section 6. Finally, Appendix A provides a graphical summary of all the analyzed interplays.

2. Background

In this section, we introduce the three identified pillars of responsible machine learning. For each of them, we briefly review their key ideas, with an emphasize on the particular aspects that will ease the understanding of subsequent sections.

2.1 Fairness

Different approaches to fairness have been proposed in the literature, which can be grouped into three main categories (Verma & Rubin, 2018). 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, ...) should not influence the predictions. *Individual fairness* approaches build on the idea that similar individuals should be treated similarly (Dwork et al., 2012). 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 (Kilbertus et al., 2017).

Depending on which step of the (supervised) ML pipeline they intervene on, fairness-enhancing methods can be divided into three main categories (Bellamy et al., 2019; Friedler et al., 2019; Caton & Haas, 2023). *Pre-processing* methods aim at removing undesired correlations from the training data before applying standard learning techniques on the sanitized data while *post-processing* techniques modify the outputs of a trained model to achieve fairness. Finally, *in-processing* (also called *algorithmic modification*) techniques directly adapt the learning procedure to produce inherently fair models.

2.2 Explainability/Interpretability

There are two main approaches towards facilitating the understanding of ML models (Burkart & Huber, 2021). On the one hand, post-hoc explanations (Guidotti et al., 2018) can be crafted to explain the behaviour of a black-box model. Depending on their form, different types of such explanations can be defined, among which *example-based explanations* consist in datapoints, belonging to the same space as the model's training set examples. For instance, they can be highly influential training examples (Koh & Liang, 2017), nearest neighbours 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 (Selvaraju et al., 2017) 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 constrast, *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 that are inherently interpretable by humans. For instance, decision trees or rule lists of reasonable size are commonly considered as interpretable (Lipton, 2018). 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.

2.3 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 (Dwork et al., 2017). More specifically in ML, the computation being performed is usually a learning algorithm whose output is a trained model. 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 (Cristofaro, 2020; Rigaki & García, 2024), among which:

- *Membership inference attacks* try to infer whether given examples were used to train a model or not (Shokri et al., 2017).
- *Reconstruction attacks* aim at reconstructing part of a model's training data (Dwork et al., 2017).
- *Model extraction attacks* aim at stealing a black-box model's internal functionalities or parameters (Tramèr et al., 2016).
- *Model inversion attacks* focus on retrieving a model's inputs by only observing the associated outputs (Fredrikson et al., 2015). 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 (Clifton & Tassa, 2013). For instance, k-anonymity (Sweeney, 2002; Samarati, 2001), requires that each block contains at least k examples. Several extensions of k-anonymity were proposed, among which t-closeness (Li et al., 2007) additionally ensures that the distribution of the values within each block 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 (Dwork et al., 2006). 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] (Abowd, 2018), but also its use by companies such as Google (Aktay et al., 2020), Facebook (Herdagdelen et al., 2020) and Apple (Team, 2017).

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 (Dwork & Roth, 2014). For instance, the *Laplace* (respectively, *Gaussian*) *mechanism* (Dwork et al., 2006) 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* (Zhang et al., 2012) 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* (McSherry & Talwar, 2007) 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 (Ji et al., 2014; Gong et al., 2020). For instance, `DP-SGD` (Abadi et al., 2016) 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 (Papernot et al., 2017, 2018). First, the (private) training set is partitioned into a number of non-overlapping subsets used to train a set of *teacher* models. Afterwards, 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 Ponomareva et al. (2023), 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. `https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/differential-privacy.html`

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. Afterwards, 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

Simplicity and fairness intrinsically conflict A framework to theoretically study the implications of enforcing interpretability is proposed by Agarwal (2021a), adapted from that of Kleinberg and Mullainathan (2019). 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 Kleinberg and Mullainathan (2019), further illustrating how simplicity can be inconsistent with statistical fairness notions. As stated by Dziugaite et al. (2020), 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.

Empirical trade-offs are complex An empirical study of the trade-offs between interpretability and fairness was conducted by Jabbari et al. (2020). 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). 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.

3.1.2 Combining Fairness and Interpretability is Challenging

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 towards interpretable machine learning (Rudin et al.,

2022). 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 (Aïvodji et al., 2022).

Explanations may not preserve fairness properties of a model It was observed by Dai et al. (2021) 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 (Aïvodji et al., 2019). On the other hand, the explanations of a fair model's decisions may (wrongly) rely on sensitive features and exhibit discrimination (Manerba & Guidotti, 2022). In addition, the choice of the explanation method as well as the type of explanation it produces both impact the users' perceived fairness (Dodge et al., 2019). The fairness of post-hoc explanations generated from a fair model's decisions was also investigated by Dai et al. (2021). 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). Afterwards, fairness is computed on a neighbourhood 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.

Fairness-enhancing methods may require non-interpretable transformations, hence harming interpretability In a study on interpretable, fair and accurate ML for criminal recidivism prediction, Wang et al. (2020) 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 (Kamiran & Calders, 2012; Zemel et al., 2013). The resulting representation hence can not be used to produce an understandable model. Furthermore, the corrections performed to a model's outputs by post-processing techniques (Pleiss et al., 2017) can also lead to non-interpretable processes.

3.1.3 Other Unfair Effects of Explainability Methods

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 Shokri et al. (2020, 2021). 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.

Post-hoc explanation frameworks can introduce unfairness through disparity in explanation quality Group-based disparities in explanation quality have been recently investigated by Dai et al. (2022). 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, Balagopalan et al. (2022) 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.

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 (Ustun et al., 2019; Sharma et al., 2020). 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 (Gupta et al., 2019; Karimi et al., 2023) and frameworks equalizing the cost of recourse across subgroups were proposed.

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 (Lakkaraju & Bastani, 2020). In addition, it has been demonstrated that this can be exploited and extended to an existing framework (Lakkaraju et al., 2019) 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 (Aïvodji et al., 2019, 2021) 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 (Slack et al., 2020). This can be leveraged to hide a black-box model's unfairness by crafting and providing fair explanations to an

auditor (Slack et al., 2021). Furthermore, Heo et al. (2019) and Dimanov et al. (2020) 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), Pruthi et al. (2020) 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 (Slack et al., 2021, 2021). In Zhang et al.'s (2020) 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, Laberge et al. (2023) 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 (Merrer & Trédan, 2019).

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

3.2 Synergies

Interpretability and explainability ease model audit As mentioned by Rudin (2019), 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, Doshi-Velez and Kim (2017) 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, Begley et al. (2020) propose to rely on *fairness explanations* based on Shapley values to be able to attribute a model's overall unfairness to individual input features.

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 (Kilbertus et al., 2018). 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 (Fioretto et al., 2022) studying the interactions between fairness and differential privacy (DP), in both decision making and machine learning tasks.

4.1 Tensions

As discussed in Section 2.1, 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, Kilbertus et al. (2018) 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.

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, Cummings et al. (2019) have shown that $(\epsilon, 0)$-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 (ϵ,δ)-DP (as δ is usually required to be cryptographically small). An impossibility theorem is also stated by Agarwal (2021b), 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 Cummings et al. (2019). (i) Consider a learning algorithm $\mathcal{L}$ that is $(\epsilon, 0)$-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 serie of datasets neighboring two-by-two, from $\mathcal{D}$ to $\mathcal{D}'$ (and the property must be verified for all pairs of neighbouring 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 meeet 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).

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 Kulynych et al. (2022). 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 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, Ekstrand et al. (2018) emphasize the importance for a privacy-preserving mechanism to protect individuals with equivalent effectiveness. However, while DP 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 Chang and Shokri (2021), 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 (Hu & Lan, 2020; Ferry et al., 2023). 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.

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 Bagdasaryan et al. (2019), 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` (Abadi et al., 2016) 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 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, Uniyal et al. (2021) empirically observe that the differentially-private `PATE` (Paper-

not et al., 2017, 2018) framework (introduced in Section 2.3) 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 privacy levels, and note that a sweet spot for the number of teachers exists, which minimizes the induced disparities. Farrand et al. (2020) observe that the accuracy disparity caused by 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 (Fioretto et al., 2022).

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 (Suriyakumar et al., 2021). 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 (Tran et al., 2021a). 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 (Xu et al., 2021). Other work (Zhang et al., 2023) 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, Tran et al. (2021a) 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. Pujol et al. (2020) 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 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 (Fioretto et al., 2022). 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 Koch and Soll (2023) 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.

Differential privacy disproportionately affects the quality of post-hoc explanations Datta et al. (2016) 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, Patel et al. (2022) 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

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 learning algorithm to output a model *approximately* satisfying given fairness criteria (Cummings et al., 2019). 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 Cummings et al. (2019). 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 Lagragian 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 Xu et al. (2019) 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 (Zhang et al., 2012). 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, Ding et al. (2020) extended `PFLR` by proposing to have two distinct privacy budgets in order 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 Tran et al. (2021b). 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. Afterwards, 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 Jagielski et al. (2019) 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 Hardt et al. (2016). 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 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 (Agarwal et al., 2018), 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 (in a similar manner as for accuracy).

Mozannar et al. (2020) 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 (Agarwal et al., 2018). Second, a modified version of a post-processing fairness-enhancing method (Hardt et al., 2016) 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 fairness violation.

The fairness cost of differential privacy can be theoretically bounded Recent work theoretically shows that the impact of DP on fairness is bounded and can be computed to obtain non-trivial guarantees regarding the private model's fairness (Mangold et al., 2023). 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 are 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.

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 (Imana et al., 2023) 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.

Individual fairness and differential privacy are both robustness definitions As introduced in Section 2.1, individual fairness can be formulated as a Lipschitz condition: just like DP, it is a robustness definition (Ignatiev et al., 2020). More precisely, Dwork et al. (2012) 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 Zemel et al. (2013), 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, Dwork et al. (2012) propose an efficient individually fair learning algorithm based on the Exponential mechanism (McSherry & Talwar, 2007), resulting in provable loss bounds. In Jagielski et al. (2019), 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 Ignatiev et al. (2020). 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.

Privacy and fairness can enhance each other in particular setups Khalili et al. (2021) 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...). They show that the Exponential mechanism can be used to perform the selection given the qualification scores, in order 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, Ruggieri (2013) 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 (Friedberg & Rogers, 2022).

Other work (Hajian et al., 2015) 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.

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

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 (Datta et al., 2023). 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 (Banisar, 2011).

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, Severi et al. (2021) 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 (Garcia et al., 2018). 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 Kuppa and Le-Khac (2021) 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.

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 (Dwork et al., 2017; Cristofaro, 2020). 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 ...), can be leveraged to enhance the different types of privacy attacks (introduced in Section 2.3):

- **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 (Milli et al., 2019). 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 (Miura et al., 2021) 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 gener-

ative model is updated leveraging the provided gradient explanations, which dramatically reduces the required number of iterations (and queries to the black-box). Furthermore, Aïvodji et al. (2020) 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 Kuppa and Le-Khac (2021), 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 Wang et al. (2022), 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 Shokri et al. (2021) 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 (Ribeiro et al., 2016; Smilkov et al., 2017) 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 (Kumar et al., 2020). 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 Kuppa and Le-Khac (2021) 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 (Koh & Liang, 2017) and returning influential training examples that most contribute to an example's prediction is considered by Shokri et al. (2021). 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.** Zhao, Zhang, Xiao, and Lim (2021) 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[4] (*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 (Simonyan et al., 2014), influence-based explanations (Ramaswamy et al., 2020) (obtained by multiplying each input feature by its associated gradient), activation-based explanations (Selvaraju et al., 2017) and layer-wise relevance propagation (Bach et al., 2015) (*i.e.*, attributing pixels' importance by back-propagating 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, Luo et al. (2022) 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

4. 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 2.3).

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 (Duddu & Boutet, 2022). 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.

Interpretable models inherently leak information regarding their training data The approach of Gambs et al. (2012) exploits the structure of a trained decision tree to reconstruct a probabilistic version of its training set. It is generalized by Ferry et al. (2024) 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.

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, Milli et al. (2019) 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, Shokri et al. (2021) 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 (Aïvodji et al., 2020).

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

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 (Doshi-Velez & Kim, 2017). It also makes it easier to detect possible privacy issues when building interpretable models (Rudin, 2019). Furthermore, this auditable nature is particularly appreciated in the area of ML-based cybersecurity systems (Srivastava et al., 2022). 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.

Interpretability can be conciliated with privacy with some trade-offs Friedman and Schuster (2010) 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 Fletcher and Islam (2019). Locally Linear Maps (LLMs) are studied by Harder et al. (2020) 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 (Abadi et al., 2016). They empirically observe a trade-off between the privacy guarantee and the model's accuracy and interpretability.

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 (Datta et al., 2016). 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 DPy 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 (Patel et al., 2022). 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. Naidu et al. (2021) investigated the impact of a model's differential privacy on the quality of post-hoc explanations (saliency maps (Selvaraju et al., 2017)) 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 2.3). 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 Severi et al. (2021) (and discussed in Section 5.1), Nguyen et al. (2023) 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 Mochaourab et al. (2021). 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, in order 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, Li et al. (2023) 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 (Yang et al., 2022). 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.

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. 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 the distribution of the noise added by privacy-preserving techniques to improve fairness (Xu et al., 2019), leveraging fairness constraints to enhance the learning of interpretable models through effective pruning mechanisms (Aïvodji et al., 2022), or leveraging explainability tools to detect privacy leakages (Srivastava et al., 2022).

Nevertheless, compromises usually have to 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. Both are summarized in Figures 1 and 2, in the Appendix A. 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. 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 are 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 Weinberg (2022) which reviews critics of popular fairness-enhancing approaches from an interdisciplinary perspective.

Appendix A. Summary Figures

In this appendix section, we provide a graphical summary of the key interplays identified between fairness, interpretability and privacy in machine learning. More precisely, we report compatibilities and synergies in Figure 1, while we overview tensions in Figure 2.

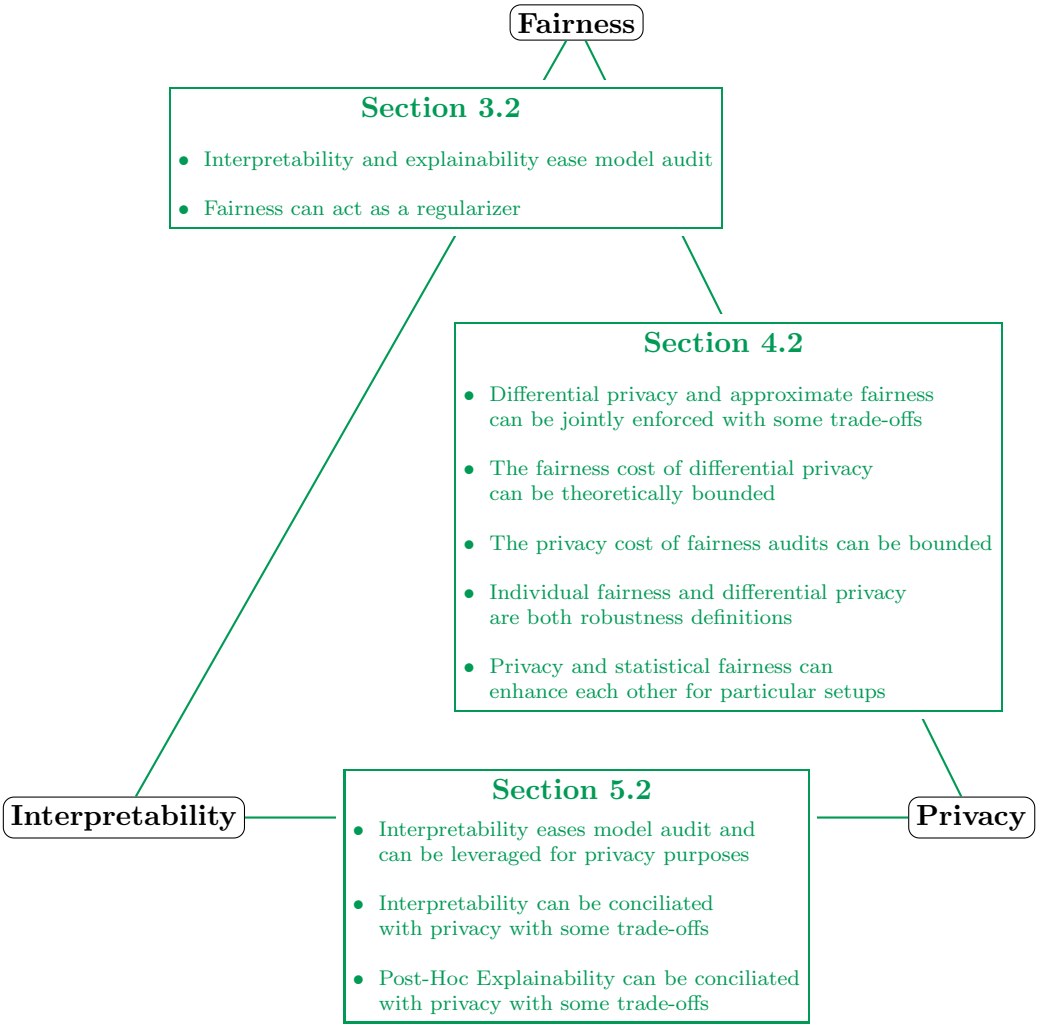

Figure 1: Summary of the identified compatibilities and synergies between fairness, interpretability and privacy in machine learning.

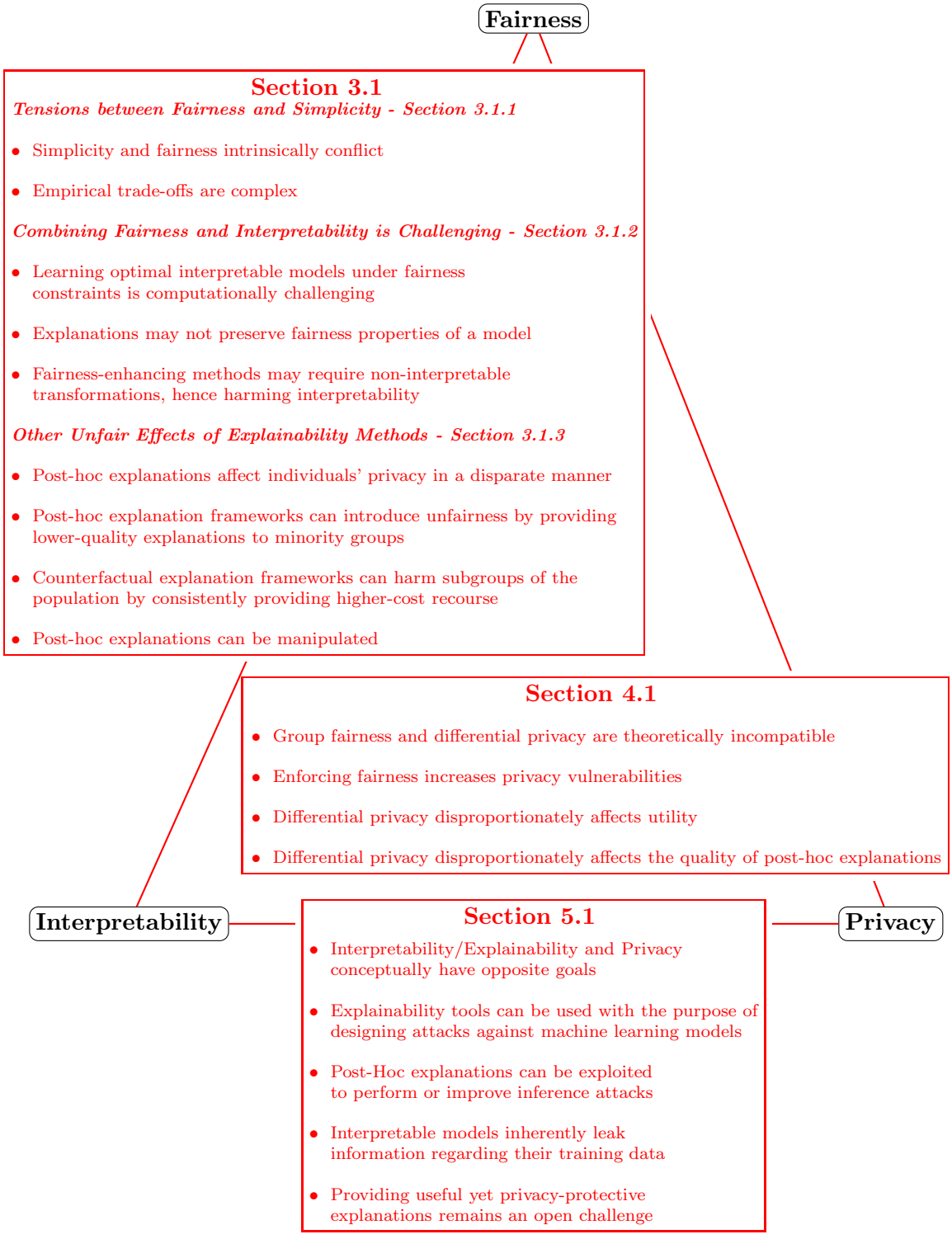

Figure 2: Summary of the identified tensions between fairness, interpretability and privacy in machine learning.

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Weippl, E. R., Katzenbeisser, S., Kruegel, C., Myers, A. C., & Halevi, S. (Eds.), *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pp. 308–318. ACM.

Abowd, J. M. (2018). The U.S. census bureau adopts differential privacy. In Guo, Y., & Farooq, F. (Eds.), *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, p. 2867. ACM.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. M. (2018). A reductions approach to fair classification. In Dy, J. G., & Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR.

Agarwal, S. (2021a). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.

Agarwal, S. (2021b). Trade-offs between fairness and privacy in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: the risk of rationalization. In Chaudhuri, K., & Salakhutdinov, R. (Eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 161–170. PMLR.

Aïvodji, U., Arai, H., Gambs, S., & Hara, S. (2021). Characterizing the risk of fairwashing. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., & Vaughan, J. W. (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 14822–14834.

Aïvodji, U., Bolot, A., & Gambs, S. (2020). Model extraction from counterfactual explanations. *CoRR*, *abs/2009.01884*.

Aïvodji, U., Ferry, J., Gambs, S., Huguet, M., & Siala, M. (2022). Leveraging integer linear programming to learn optimal fair rule lists. In Schaus, P. (Ed.), *Integration of Constraint Programming, Artificial Intelligence, and Operations Research - 19th International Conference, CPAIOR 2022, Los Angeles, CA, USA, June 20-23, 2022, Proceedings*, Vol. 13292 of *Lecture Notes in Computer Science*, pp. 103–119. Springer.

Aktay, A., Bavadekar, S., Cossoul, G., Davis, J., Desfontaines, D., Fabrikant, A., Gabrilovich, E., Gadepalli, K., Gipson, B., Guevara, M., Kamath, C., Kansal, M., Lange, A., Mandayam, C., Oplinger, A., Pluntke, C., Roessler, T., Schlosberg, A., Shekel, T., Vispute, S., Vu, M., Wellenius, G., Williams, B., & Wilson, R. J. (2020). Google COVID-19 community mobility reports: Anonymization process description (version 1.0). *CoRR*, *abs/2004.04145*.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), e0130140.

Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 15453–15462.

Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., & Ghassemi, M. (2022). The road to explainability is paved with bias: Measuring the fairness of explanations. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pp. 1194–1206. ACM.

Banisar, D. (2011). The right to information and privacy: balancing rights and managing conflicts..

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. `http://www.fairmlbook.org`.

Begley, T., Schwedes, T., Frye, C., & Feige, I. (2020). Explainability for fair machine learning. *CoRR*, *abs/2010.07389*.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.*, *63*(4/5), 4:1–4:15.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.*, *70*, 245–317.

Caton, S., & Haas, C. (2023). Fairness in machine learning: A survey.. New York, NY, USA. Association for Computing Machinery.

Chang, H., & Shokri, R. (2021). On the privacy risks of algorithmic fairness. In *IEEE European Symposium on Security and Privacy, EuroS&P 2021, Vienna, Austria, September 6-10, 2021*, pp. 292–303. IEEE.

Clifton, C., & Tassa, T. (2013). On syntactic anonymity and differential privacy. *Trans. Data Priv.*, *6*(2), 161–183.

Cristofaro, E. D. (2020). An overview of privacy in machine learning. *CoRR*, *abs/2005.08679*.

Cummings, R., Gupta, V., Kimpara, D., & Morgenstern, J. (2019). On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, UMAP'19 Adjunct, p. 309–315, New York, NY, USA. Association for Computing Machinery.

Dai, J., Upadhyay, S., Aïvodji, U., Bach, S. H., & Lakkaraju, H. (2022). Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In Conitzer, V., Tasioulas, J., Scheutz, M., Calo, R., Mara, M., & Zimmermann, A.

(Eds.), *AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, May 19 - 21, 2021*, pp. 203–214. ACM.

Dai, J., Upadhyay, S., Bach, S. H., & Lakkaraju, H. (2021). What will it take to generate fairness-preserving explanations?. *CoRR*, *abs/2106.13346*.

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pp. 598–617. IEEE Computer Society.

Datta, T., Nissani, D., Cembalest, M., Khanna, A., Massa, H., & Dickerson, J. P. (2023). Position: Tensions between the proxies of human values in AI. In *First IEEE Conference on Secure and Trustworthy Machine Learning*.

Dimanov, B., Bhatt, U., Jamnik, M., & Weller, A. (2020). You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In Giacomo, G. D., Catalá, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., & Lang, J. (Eds.), *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, Vol. 325 of *Frontiers in Artificial Intelligence and Applications*, pp. 2473–2480. IOS Press.

Ding, J., Zhang, X., Li, X., Wang, J., Yu, R., & Pan, M. (2020). Differentially private and fair classification via calibrated functional mechanism. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 622–629. AAAI Press.

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019). Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*, pp. 275–285.

Doshi-Velez, F., & Kim, B. (2017). A roadmap for a rigorous science of interpretability. *CoRR*, *abs/1702.08608*.

Duddu, V., & Boutet, A. (2022). Inferring sensitive attributes from model explanations. In Hasan, M. A., & Xiong, L. (Eds.), *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pp. 416–425. ACM.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, p. 214–226, New York, NY, USA. Association for Computing Machinery.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. D. (2006). Calibrating noise to sensitivity in private data analysis. In Halevi, S., & Rabin, T. (Eds.), *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March*

4-7, 2006, Proceedings, Vol. 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, *9*(3–4), 211–407.

Dwork, C., Smith, A., Steinke, T., & Ullman, J. (2017). Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, *4*(1), 61–84.

Dziugaite, G. K., Ben-David, S., & Roy, D. M. (2020). Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *CoRR*, *abs/2010.13764*.

Ekstrand, M. D., Joshaghani, R., & Mehrpouyan, H. (2018). Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*, pp. 35–47. PMLR.

Farrand, T., Mireshghallah, F., Singh, S., & Trask, A. (2020). Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In Zhang, B., Popa, R. A., Zaharia, M., Gu, G., & Ji, S. (Eds.), *PPMLP'20: Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice, Virtual Event, USA, November, 2020*, pp. 15–19. ACM.

Ferry, J., Aïvodji, U., Gambs, S., Huguet, M.-J., & Siala, M. (2023). Exploiting fairness to enhance sensitive attributes reconstruction. In *First IEEE Conference on Secure and Trustworthy Machine Learning*.

Ferry, J., Aïvodji, U., Gambs, S., Huguet, M.-J., & Siala, M. (2024). Probabilistic Dataset Reconstruction from Interpretable Models. In *2nd IEEE Conference on Secure and Trustworthy Machine Learning*, Toronto, Canada.

Fioretto, F., Tran, C., Hentenryck, P. V., & Zhu, K. (2022). Differential privacy and fairness in decisions and learning tasks: A survey. In Raedt, L. D. (Ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pp. 5470–5477. ijcai.org.

Fletcher, S., & Islam, M. Z. (2019). Decision tree classification with differential privacy: A survey. *ACM Comput. Surv.*, *52*(4), 83:1–83:33.

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Ray, I., Li, N., & Kruegel, C. (Eds.), *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015*, pp. 1322–1333. ACM.

Friedberg, R., & Rogers, R. (2022). Privacy aware experimentation over sensitive groups: A general chi square approach. In Dieng, A., Rateike, M., Farnadi, G., Fioretto, F., Kusner, M. J., & Schrouff, J. (Eds.), *Algorithmic Fairness through the Lens of Causality and Privacy Workshop, AFCP 2022, New Orleans, LA, USA (hybrid), 03 December 2022*, Vol. 214 of *Proceedings of Machine Learning Research*, pp. 23–66. PMLR.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine

learning. In danah boyd, & Morgenstern, J. H. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 329–338. ACM.

Friedman, A., & Schuster, A. (2010). Data mining with differential privacy. In Rao, B., Krishnapuram, B., Tomkins, A., & Yang, Q. (Eds.), *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pp. 493–502. ACM.

Gambs, S., Gmati, A., & Hurfin, M. (2012). Reconstruction attack through classifier analysis. In Cuppens-Boulahia, N., Cuppens, F., & García-Alfaro, J. (Eds.), *Data and Applications Security and Privacy XXVI - 26th Annual IFIP WG 11.3 Conference, DBSec 2012, Paris, France, July 11-13,2012. Proceedings*, Vol. 7371 of *Lecture Notes in Computer Science*, pp. 274–281. Springer.

Garcia, W., Choi, J. I., Adari, S. K., Jha, S., & Butler, K. R. B. (2018). Explainable black-box attacks against model-based authentication. *CoRR*, *abs/1810.00024*.

Gong, M., Xie, Y., Pan, K., Feng, K., & Qin, A. K. (2020). A survey on differentially private machine learning [review article]. *IEEE Comput. Intell. Mag.*, *15*(2), 49–64.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42.

Gupta, V., Nokhiz, P., Roy, C. D., & Venkatasubramanian, S. (2019). Equalizing recourse across groups. *CoRR*, *abs/1909.03166*.

Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., & Giannotti, F. (2015). Discrimination- and privacy-aware patterns. *Data Min. Knowl. Discov.*, *29*(6), 1733–1782.

Harder, F., Bauer, M., & Park, M. (2020). Interpretable and differentially private predictions. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 4083–4090. AAAI Press.

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc.

Heo, J., Joo, S., & Moon, T. (2019). Fooling neural network interpretations via adversarial model manipulation. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 2921–2932.

Herdagdelen, A., Dow, A., State, B., Mohassel, P., & Pompe, A. (2020). Protecting privacy in facebook mobility data during the covid- 19 response..

Hu, H., & Lan, C. (2020). Inference attack and defense on the distributed private fair learning framework. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence*.

Ignatiev, A., Cooper, M. C., Siala, M., Hebrard, E., & Marques-Silva, J. (2020). Towards formal fairness in machine learning. In *International Conference on Principles and Practice of Constraint Programming*, pp. 846–867. Springer.

Imana, B., Korolova, A., & Heidemann, J. S. (2023). Having your privacy cake and eating it too: Platform-supported auditing of social media algorithms for public interest. *Proc. ACM Hum. Comput. Interact.*, *7*(CSCW1), 1–33.

Jabbari, S., Ou, H.-C., Lakkaraju, H., & Tambe, M. (2020). An empirical study of the trade-offs between interpretability and fairness. In *ICML 2020 Workshop on Human Interpretability in Machine Learning*.

Jagielski, M., Kearns, M. J., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., & Ullman, J. R. (2019). Differentially private fair learning. In Chaudhuri, K., & Salakhutdinov, R. (Eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 3000–3008. PMLR.

Ji, Z., Lipton, Z. C., & Elkan, C. (2014). Differential privacy and machine learning: a survey and review. *CoRR*, *abs/1412.7584*.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33.

Karimi, A., Barthe, G., Schölkopf, B., & Valera, I. (2023). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput. Surv.*, *55*(5), 95:1–95:29.

Khalili, M. M., Zhang, X., Abroshan, M., & Sojoudi, S. (2021). Improving fairness and privacy in selection problems. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 8092–8100. AAAI Press.

Kilbertus, N., Gascón, A., Kusner, M. J., Veale, M., Gummadi, K. P., & Weller, A. (2018). Blind justice: Fairness with encrypted sensitive attributes. In Dy, J. G., & Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 2635–2644. PMLR.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 656–666.

Kleinberg, J., & Mullainathan, S. (2019). Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 807–808.

Koch, K., & Soll, M. (2023). 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*.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR.

Kulynych, B., Yaghini, M., Cherubin, G., Veale, M., & Troncoso, C. (2022). Disparate vulnerability to membership inference attacks. *Proc. Priv. Enhancing Technol.*, *2022*(1), 460–480.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491–5500. PMLR.

Kuppa, A., & Le-Khac, N. (2021). Adversarial XAI methods in cybersecurity. *IEEE Trans. Inf. Forensics Secur.*, *16*, 4924–4938.

Laberge, G., Aïvodji, U., Hara, S., Marchand, M., & Khomh, F. (2023). Fooling SHAP with stealthily biased sampling. In *The Eleventh International Conference on Learning Representations*.

Lakkaraju, H., & Bastani, O. (2020). "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 '20, p. 79–85, New York, NY, USA. Association for Computing Machinery.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In Conitzer, V., Hadfield, G. K., & Vallor, S. (Eds.), *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pp. 131–138. ACM.

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In Chirkova, R., Dogac, A., Özsu, M. T., & Sellis, T. K. (Eds.), *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pp. 106–115. IEEE Computer Society.

Li, Z., Chen, H., Ni, Z., & Shao, H. (2023). Balancing privacy protection and interpretability in federated learning. *CoRR*, *abs/2302.08044*.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31–57.

Luo, X., Jiang, Y., & Xiao, X. (2022). Feature inference attack on shapley values. In Yin, H., Stavrou, A., Cremers, C., & Shi, E. (Eds.), *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, pp. 2233–2247. ACM.

Manerba, M. M., & Guidotti, R. (2022). Investigating debiasing effects on classification and explainability. In Conitzer, V., Tasioulas, J., Scheutz, M., Calo, R., Mara, M., & Zimmermann, A. (Eds.), *AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, May 19 - 21, 2021*, pp. 468–478. ACM.

Mangold, P., Perrot, M., Bellet, A., & Tommasi, M. (2023). Differential privacy has bounded impact on fairness in classification. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., & Scarlett, J. (Eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 23681–23705. PMLR.

McSherry, F., & Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, pp. 94–103. IEEE Computer Society.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, *54*(6), 115:1–115:35.

Merrer, E. L., & Trédan, G. (2019). The bouncer problem: Challenges to remote explainability. *CoRR*, *abs/1910.01432*.

Milli, S., Schmidt, L., Dragan, A. D., & Hardt, M. (2019). Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 1–9.

Miura, T., Hasegawa, S., & Shibahara, T. (2021). MEGEX: data-free model extraction attack against gradient-based explainable AI. *CoRR*, *abs/2107.08909*.

Mochaourab, R., Sinha, S., Greenstein, S., & Papapetrou, P. (2021). Robust explanations for private support vector machines. *CoRR*, *abs/2102.03785*.

Mozannar, H., Ohannessian, M., & Srebro, N. (2020). Fair learning with private demographic data. In *International Conference on Machine Learning*, pp. 7066–7075. PMLR.

Naidu, R., Priyanshu, A., Kumar, A., Kotti, S., Wang, H., & Mireshghallah, F. (2021). When differential privacy meets interpretability: A case study. *CoRR*, *abs/2106.13203*.

Nguyen, T. D. T., Lai, P., Phan, H., & Thai, M. T. (2023). Xrand: Differentially private defense against explanation-guided attacks. In Williams, B., Chen, Y., & Neville, J. (Eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 11873–11881. AAAI Press.

Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I. J., & Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., & Erlingsson, Ú. (2018). Scalable private learning with PATE. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Patel, N., Shokri, R., & Zick, Y. (2022). Model explanations with differential privacy. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pp. 1895–1904. ACM.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, *30*.

Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H. B., Vassilvitskii, S., Chien, S., & Thakurta, A. G. (2023). How to dp-fy ML: A practical guide to machine learning with differential privacy. *J. Artif. Intell. Res.*, *77*, 1113–1201.

Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., & Lipton, Z. C. (2020). Learning to deceive with attention-based explanations. In Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. R. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4782–4793. Association for Computational Linguistics.

Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., & Miklau, G. (2020). Fair decision making using privacy-protected data. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 189–199. Association for Computing Machinery, Inc.

Ramaswamy, H. G., et al. (2020). 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*, pp. 983–991.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " 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*, pp. 1135–1144.

Rigaki, M., & García, S. (2024). A survey of privacy attacks in machine learning. *ACM Comput. Surv.*, *56*(4), 101:1–101:34.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, *16*, 1–85.

Ruggieri, S. (2013). Data anonymity meets non-discrimination. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pp. 875–882. IEEE.

Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.*, *13*(6), 1010–1027.

Schöffer, J. (2023). *On the Interplay of Transparency and Fairness in AI-Informed Decision-Making*. Ph.D. thesis, Karlsruher Institut für Technologie (KIT).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Severi, G., Meyer, J., Coull, S. E., & Oprea, A. (2021). Explanation-guided backdoor poisoning attacks against malware classifiers. In Bailey, M., & Greenstadt, R. (Eds.), *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pp. 1487–1504. USENIX Association.

Sharma, S., Henderson, J., & Ghosh, J. (2020). CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In Markham, A. N., Powles, J., Walsh, T., & Washington, A. L. (Eds.), *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, pp. 166–172. ACM.

Shokri, R., Strobel, M., & Zick, Y. (2020). Exploiting transparency measures for membership inference: a cautionary tale. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI). AAAI*, Vol. 13.

Shokri, R., Strobel, M., & Zick, Y. (2021). On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 231–241.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 3–18. IEEE Computer Society.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In Bengio, Y., & LeCun, Y. (Eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Slack, D., Hilgard, A., Lakkaraju, H., & Singh, S. (2021). Counterfactual explanations can be manipulated. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., & Vaughan, J. W. (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 62–75.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, p. 180–186, New York, NY, USA. Association for Computing Machinery.

Slack, D., Hilgard, S., Singh, S., & Lakkaraju, H. (2021). Feature attributions and counterfactual explanations can be manipulated. *CoRR*, *abs/2106.12563*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *CoRR*, *abs/1706.03825*.

Srivastava, G., Jhaveri, R. H., Bhattacharya, S., Pandya, S., Rajeswari, Maddikunta, P. K. R., Yenduri, G., Hall, J. G., Alazab, M., & Gadekallu, T. R. (2022). XAI for cybersecurity: State of the art, challenges, open issues and future directions. *CoRR*, *abs/2206.03585*.

Suriyakumar, V. M., Papernot, N., Goldenberg, A., & Ghassemi, M. (2021). Chasing your long tails: Differentially private prediction in health care settings. In Elish, M. C., Isaac, W., & Zemel, R. S. (Eds.), *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pp. 723–734. ACM.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, *10*(5), 557–570.

Team, A. D. P. (2017). Learning with privacy at scale..

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction apis. In Holz, T., & Savage, S. (Eds.), *25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016*, pp. 601–618. USENIX Association.

Tran, C., Dinh, M. H., Beiter, K., & Fioretto, F. (2021a). A fairness analysis on private aggregation of teacher ensembles. *CoRR*, *abs/2109.08630*.

Tran, C., Fioretto, F., & Hentenryck, P. V. (2021b). Differentially private and fair deep learning: A lagrangian dual approach. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 9932–9939. AAAI Press.

Uniyal, A., Naidu, R., Kotti, S., Singh, S., Kenfack, P. J., Mireshghallah, F., & Trask, A. (2021). DP-SGD vs PATE: which has less disparate impact on model accuracy?. *CoRR*, *abs/2106.12576*.

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In danah boyd, & Morgenstern, J. H. (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 10–19. ACM.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In Brun, Y., Johnson, B., & Meliou, A. (Eds.), *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, pp. 1–7. ACM.

Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, *10*(3152676), 10–5555.

Wang, C., Han, B., Patel, B., Mohideen, F., & Rudin, C. (2020). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *CoRR*, *abs/2005.04176*.

Wang, Y., Qian, H., & Miao, C. (2022). Dualcf: Efficient model extraction attack from counterfactual explanations. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pp. 1318–1329. ACM.

Weinberg, L. (2022). Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches. *J. Artif. Intell. Res.*, *74*, 75–109.

Xu, D., Du, W., & Wu, X. (2021). Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In Zhu, F., Ooi, B. C., & Miao, C. (Eds.), *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pp. 1924–1932. ACM.

Xu, D., Yuan, S., & Wu, X. (2019). Achieving differential privacy and fairness in logistic regression. In Amer-Yahia, S., Mahdian, M., Goel, A., Houben, G., Lerman, K.,

McAuley, J. J., Baeza-Yates, R., & Zia, L. (Eds.), *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pp. 594–599. ACM.

Yang, F., Feng, Q., Zhou, K., Chen, J., & Hu, X. (2022). Differentially private counterfactuals via functional mechanism. *CoRR*, *abs/2208.02878*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333.

Zhang, J., Zhang, Z., Xiao, X., Yang, Y., & Winslett, M. (2012). Functional mechanism: Regression analysis under differential privacy. *Proc. VLDB Endow.*, *5*(11), 1364–1375.

Zhang, T., Zhu, T., Gao, K., Zhou, W., & Yu, P. S. (2023). Balancing learning model privacy, fairness, and accuracy with early stopping criteria. *IEEE Trans. Neural Networks Learn. Syst.*, *34*(9), 5557–5569.

Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., & Wang, T. (2020). Interpretable deep learning under fire. In Capkun, S., & Roesner, F. (Eds.), *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, pp. 1659–1676. USENIX Association.

Zhao, X., Zhang, W., Xiao, X., & Lim, B. (2021). Exploiting explanations for model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 682–692.