Survey on Fairness Notions and Related Tensions

Guilherme Alves, Fabien Bernier, Miguel Couceiro, Karima Makhlouf,
Catuscia Palamidessi, Sami Zhioua

Noname manuscript No.
(will be inserted by the editor)

Survey on Fairness Notions and Related Tensions.

Guilherme Alves · Fabien Bernier · Miguel Couceiro · Karima Makhlouf · Catuscia Palamidessi · Sami Zhioua

Received: date / Accepted: date

Abstract Automated decision systems are increasingly used to take consequential decisions in problems such as job hiring and loan granting with the hope of replacing subjective human decisions with objective machine learning (ML) algorithms. ML-based decision systems, however, are found to be prone to bias which result in yet unfair decisions. Several notions of fairness have been defined in the literature to capture the different subtleties of this ethical and social concept (e.g. statistical parity, equal opportunity, etc.). Fairness requirements to be satisfied while learning models created several types of tensions among the different notions of fairness, but also with other desirable properties such as privacy and classification accuracy. This paper surveys the commonly used fairness notions and discusses the tensions that exist among them and with privacy and accuracy. It also shows how the simple idea of fairness through unawareness (dropping sensitive features) can be leveraged through explanations and ensemble learning to appropriately address the tension between fairness and classification accuracy.

Keywords Fairness notion · Tension within fairness · Explanation method · Unfairness mitigation

Author ordering on this paper is alphabetical.
The research of the first three named authors was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215, and the Inria Project Lab "Hybrid Approaches for Interpretable AI" (HyAIAI).
The research of the last three named authors was supported by HYPATIA, a project funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under GA No 835294.

Guilherme Alves, Fabien Bernier, Miguel Couceiro
Université de Lorraine, CNRS, Inria N.G.E., LORIA, F-54000 Nancy, France
E-mail: {guilherme.alves-da-silva, fabien.bernier, miguel.couceiro}@loria.fr

Karima Makhlouf, Catuscia Palamidessi, Sami Zhioua
Inria, École Polytechnique, IPP, 91120, Paris, France
E-mail: {karima.makhlouf, catuscia.palamidessi, sami.zhioua}@inria.fr

1 Introduction

Fairness emerged as an important requirement to guarantee that machine learning (ML) based decision systems can be safely used in practice. Using such systems while fairness is not satisfied can lead to unfair decisions typically discriminating against disadvantaged populations such as racial minorities, women, poverty stricken districts, etc.

With the recent interest for fairness, a multitude of fairness notions have been defined to capture different aspects of fairness. These include statistical group-based notions (e.g. statistical parity [19], equalized odds [26], etc.), individual-based notions (e.g. fairness through awareness [19], etc.), and causal-based notions (e.g. total effect [36], counterfactual fairness [31], etc.). As fairness is a social and ethical concept in the first place, defining it is still prone to subjectivity. Hence, the aim of replacing the subjective human decisions by objective ML-based decision systems resulted in notions and algorithms still exhibiting unfairness. Hence although the different notions of algorithmic fairness appear internally consistent, several of them cannot hold simultaneously and hence are mutually incompatible [5,35,26,6,30]. As a consequence, practitioners assessing and/or implementing fairness need to choose among them.

In addition to the tensions "intra-notions", there are tensions between fairness notions and other desirable properties of ML algorithms. One such properties is privacy. This property emerged as a concern that specific information about an individual might be revealed when a model is learned based on a dataset containing that individual. A learning algorithm satisfying privacy will learn aggregate information about the population, but will not incorporate specific information about individuals. Recent results showed that fairness and privacy (differential privacy [18]) are at odds with each other [14,2,38]. That is, a learning algorithm that satisfies differential privacy is not guaranteed to generate a classifier that satisfies fairness, unless it has trivial accuracy.

Fairness through unawareness is one of the simplest approaches to address fairness and consists in dropping the sensitive feature before training the ML model. Such approach is an example of *process fairness* [24] which focuses on the fairness of the learning process rather than the fairness of the output. Dropping features, however, creates a tension with classification accuracy: it typically improves fairness but reduces accuracy. An interesting line of research in the literature consists in handling this fairness/accuracy trade-off by aggregating multiple classifier's outputs (ensemble classifier) each of which is obtained by dropping one single sensitive feature [3]. Selecting sensitive features to drop is based on explanation methods which assess the contribution of each sensitive feature on the outcome. Examples of explanation methods include LIME [23,39] and SHAP [33].

Several surveys of the relatively recent field of ML fairness can be found in the literature [6,34,46,35,49,22]. However, this survey deviates from existing surveys by focusing on the tensions that exist among fairness notions and between fairness and other desirable ML properties, namely, privacy and classi-

fication accuracy. Section 2 briefly presents all commonly used fairness notions spanning all categories (group, individual, and causal) along with their formal definitions. Section 3 describes the tensions and incompatibilities that exist among the various fairness notions. Section 4 shows that fairness and privacy properties are at odds with each other and present a complete formal proof of this incompatibility. Section 5 describes how fairness through unawareness can be pushed further to address the fairness vs accuracy trade-off and obtain a fair but also accurate classifier thanks to process fairness, explanation methods, and ensemble classifiers. Section 6 shows how fairness notions can applied on benchmark and real datasets and illustrates some of the tensions described in the previous sections. Section 7 discusses directions for future work.

2 Catalogue of fairness notions

Let V, A, and X[1] be three random variables representing, respectively, the total set of features, the sensitive features, and the remaining features describing an individual such that $V = (X, A)$ and $P(V = v_i)$ represents the probability of drawing an individual with a vector of values v_i from the population. For simplicity, we focus on the case where A is a binary random variable where $A = 0$ designates the protected group, while $A = 1$ designates the non-protected group. Let Y represent the actual outcome and $\hat{Y}$ represent the outcome returned by the prediction algorithm. Without loss of generality, assume that Y and $\hat{Y}$ are binary random variables where $Y = 1$ designates a positive instance, while $Y = 0$ a negative one. Typically, the predicted outcome $\hat{Y}$ is derived from a score represented by a random variable S where $\mathbb{P}[S = s]$ is the probability that the score value is equal to s.

A simple and straightforward approach to address fairness problem is to ignore completely any sensitive feature while training the prediction system. This is called **fairness through unawareness**[2]. This notion is investigated further in Section 5.

Statistical parity[19] is one of the most commonly accepted notions of fairness. It requires the prediction to be statistically independent of the sensitive feature ($\hat{Y} \perp A$). In other words, the predicted acceptance rates for both protected and unprotected groups should be equal. Statistical parity implies that

$$\frac{TP + FP}{TP + FP + FN + TN}$$[3]

is equal for both groups. A classifier $\hat{Y}$ satisfies statistical parity if:

$$\mathbb{P}[\hat{Y} \mid A = 0] = \mathbb{P}[\hat{Y} \mid A = 1]. \quad (1)$$

[1] Table 10 in the appendix lists all terms used in this survey.

[2] Known also as: blindness, unawareness [35], anti-classification [11], and treatment parity [32].

[3] TP, FP, FN, and TN stand for: true positives, false positives, false negatives, and true negatives, respectively.

Conditional statistical parity [12] is a variant of statistical parity obtained by controlling on a set of resolving features[4]. The resolving features (we refer to them as R) among X are correlated with the sensitive feature A and give some factual information about the label at the same time leading to a *legitimate* discrimination. Conditional statistical parity holds if:

$$\mathbb{P}[\hat{Y} = 1 \mid R = r, A = 0] = \mathbb{P}[\hat{Y} = 1 \mid R = r, A = 1] \quad \forall r \in range(R). \tag{2}$$

Equalized odds [26] considers both the predicted and the actual outcomes. The prediction is conditionally independent from the protected feature, given the actual outcome ($\hat{Y} \perp A \mid Y$). In other words, equalized odds requires that both sub-populations to have the same true positive rate $TPR = \frac{TP}{TP+FN}$ and false positive rate $FPR = \frac{FP}{FP+TN}$:

$$\mathbb{P}[\hat{Y} = 1 \mid Y = y,\ A = 0] = \mathbb{P}[\hat{Y} = 1 \mid Y = y,\ A = 1] \quad \forall y \in \{0, 1\}. \tag{3}$$

Because equalized odds requirement is rarely satisfied in practice, two variants can be obtained by relaxing its equation. The first one is called **equal opportunity** [26] and is obtained by requiring only TPR equality among groups:

$$\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 0] = \mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 1]. \tag{4}$$

As TPR does not take into consideration FP, equal opportunity is completely insensitive to the number of false positives.

The second relaxed variant of equalized odds is called **predictive equality** [12] which requires only the FPR to be equal in both groups:

$$\mathbb{P}[\hat{Y} = 1 \mid Y = 0, A = 0] = \mathbb{P}[\hat{Y} = 1 \mid Y = 0, A = 1]. \tag{5}$$

Since FPR is independent from FN, predictive equality is completely insensitive to false negatives.

Conditional use accuracy equality [6] is achieved when all population groups have equal positive predictive value $PPV = \frac{TP}{TP+FP}$ and negative predictive value $NPV = \frac{TN}{FN+TN}$. In other words, the probability of subjects with positive predictive value to truly belong to the positive class and the probability of subjects with negative predictive value to truly belong to the negative class should be the same. By contrast to equalized odds, one is conditioning on the algorithm's predicted outcome not the actual outcome. In other words, the emphasis is on the precision of prediction rather than its recall:

$$\mathbb{P}[Y = y \mid \hat{Y} = y, A = 0] = \mathbb{P}[Y = y \mid \hat{Y} = y, A = 1] \quad \forall y \in \{0, 1\}. \tag{6}$$

Predictive parity [10] is a relaxation of conditional use accuracy equality requiring only equal PPV among groups:

$$\mathbb{P}[Y = 1 \mid \hat{Y} = 1, A = 0] = \mathbb{P}[Y = 1 \mid \hat{Y} = 1, A = 1] \tag{7}$$

Like predictive equality, predictive parity is insensitive to false negatives.

[4] Called explanatory features in [27].

Overall accuracy equality [6] is achieved when overall accuracy for both groups is the same. This implies that

$$\frac{TP + TN}{TP + FN + FP + TN}$$

is equal for both groups:

$$\mathbb{P}[\hat{Y} = Y|A = 0] = \mathbb{P}[\hat{Y} = Y|A = 1] \tag{8}$$

Treatment equality [6] is achieved when the ratio of FPs and FNs is the same for both protected and unprotected groups:

$$\frac{FN}{FP}^{\mathrm{A=0}} = \frac{FN}{FP}^{\mathrm{A=1}} \tag{9}$$

Total fairness [6] holds when all aforementioned fairness notions are satisfied simultaneously, that is, statistical parity, equalized odds, conditional use accuracy equality (hence, overall accuracy equality), and treatment equality. Total fairness is a very strong notion which is very difficult to hold in practice.

Balance [30] uses the score (S) from which the outcome Y is typically derived through thresholding. **Balance for positive class** focuses on the applicants who constitute positive instances and is satisfied if the average score S received by those applicants is the same for both groups:

$$E[S \mid Y = 1, A = 0)] = E[S \mid Y = 1, A = 1]. \tag{10}$$

Balance of negative class focuses instead on the negative class:

$$E[S \mid Y = 0, A = 0] = E[S \mid Y = 0, A = 1]. \tag{11}$$

Calibration [10] holds if, for each predicted probability score $S = s$, individuals in all groups have the same probability to actually belong to the positive class:

$$\mathbb{P}[Y = 1 \mid S = s, A = 0] = \mathbb{P}[Y = 1 \mid S = s, A = 1] \quad \forall s \in [0, 1]. \tag{12}$$

Well-calibration [30] is a stronger variant of calibration. It requires that (1) calibration is satisfied, (2) the score is interpreted as the probability to truly belong to the positive class, and (3) for each score $S = s$, the probability to truly belong to the positive class is equal to that particular score:

$$\mathbb{P}[Y = 1 \mid S = s, A = 0] = \mathbb{P}[Y = 1 \mid S = s, A = 1] = s \quad \forall\, s \in [0, 1]. \tag{13}$$

Fairness through awareness [19] implies that similar individuals should have similar predictions. Let i and j be two individuals represented by their attributes values vectors v_i and v_j. Let $d(v_i, v_j)$ represent the similarity distance between individuals i and j. Let $M(v_i)$ represent the probability distribution over the outcomes of the prediction. For example, if the outcome is binary (0 or 1), $M(v_i)$ might be $[0.2, 0.8]$ which means that for individual i, $\mathbb{P}[\hat{Y} = 0]) = 0.2$

and $\mathbb{P}[\hat{Y} = 1] = 0.8$. Let d_M be a distance metric between probability distributions. Fairness through awareness is achieved iff, for any pair of individuals i and j:

$$d_M(M(v_i), M(v_j)) \leq d(v_i, v_j)$$

In practice, fairness through awareness assumes that the similarity metric is known for each pair of individuals [29]. That is, a challenging aspect of this approach is the difficulty to determine what is an appropriate metric function to measure the similarity between two individuals. Typically, this requires careful human intervention from professionals with domain expertise [31].

Causality-based fairness notions differ from all statistical fairness approaches because they are not totally based on data but consider additional knowledge about the structure of the world, in the form of a causal model. Therefore, most of these fairness notions are defined in terms of non-observable quantities such as interventions (to simulate random experiments) and counterfactuals (which consider other hypothetical worlds, in addition to the actual world).

A variable X is a *cause* of a variable Y if Y in any way relies on X for its value [37]. Causal relationships are expressed using structural equations [8] and represented by causal graphs where nodes represent variables (features) and edges represent causal relationships between variables. Figure 1 shows a possible causal graph for the job hiring example where directed edges indicate causal relationships.

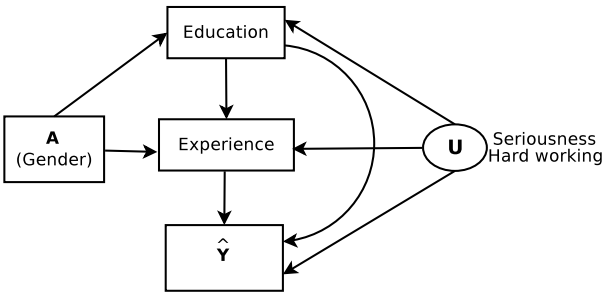


Fig. 1: A possible causal graph for the hiring example.

Total effect (TE) [36] is the causal version of statistical parity and is defined in terms of experimental probabilities as follows:

$$TE_{a_1,a_0}(\hat{y}) = \mathbb{P}[\hat{y}_{A \leftarrow a_1}] - \mathbb{P}[\hat{y}_{A \leftarrow a_0}] \tag{14}$$

where $\mathbb{P}[\hat{y}_{A \leftarrow a}] = \mathbb{P}[\hat{Y} = \hat{y} \mid do(A = a)]$ is called the experimental probability and is expressed using intervention. An intervention, denoted $do(V = v)$, is a manipulation of the model that consists in fixing the value of a variable (or a set of variables) to a specific value. Graphically, it consists in discarding all edges incident to the vertex corresponding to variable V. Intuitively, using the

job hiring example, while $\mathbb{P}[\hat{Y} = 1 \mid A = 0]$ reflects the probability of hiring among female applicants, $\mathbb{P}[\hat{Y}_{A\leftarrow 0} = 1] = \mathbb{P}[\hat{Y} = 1) \mid do(A = 0)]$ reflects the probability of hiring if *all the candidates in the population* had been female. The obtained distribution $\mathbb{P}[\hat{Y}_{A\leftarrow a}]$ can be considered as a *counterfactual* distribution since the intervention forces A to take a value different from the one it would take in the actual world. Such counterfactual variable is also denoted $\hat{Y}_{A=a}$ or $\hat{Y}_a$ for short.

TE measures the effect of the change of A from a_1 to a_0 on $\hat{Y} = \hat{y}$ along all the causal paths from A to $\hat{Y}$. Intuitively, while statistical parity reflects the difference in proportions of $\hat{Y} = \hat{y}$ in the current cohort, TE reflects the difference in proportions of $\hat{Y} = \hat{y}$ in the entire population. A more involved causal-based fairness notion considers the effect of a change in the sensitive feature value (e.g. gender) on the outcome (e.g. probability of hiring) given that we already observed the outcome for that individual. This typically involves an impossible situation which requires to go back in the past and change the sensitive feature value. Mathematically, this can be formalized using counterfactual quantities. The simplest fairness notion using counterfactuals is **the effect of treatment on the treated (ETT)** [36] defined as:

$$ETT_{a_1,a_0}(\hat{y}) = \mathbb{P}[\hat{y}_{A\leftarrow a_1} \mid a_0] - \mathbb{P}[\hat{y} \mid a_0] \tag{15}$$

$\mathbb{P}[\hat{y}_{A\leftarrow a_1} \mid a_0]$ reads the probability of $\hat{Y} = \hat{y}$ had A been a_1, given A had been observed to be a_0. For instance, in the job hiring example, $\mathbb{P}[\hat{Y}_{A\leftarrow 1} \mid A = 0]$ reads the probability of hiring an applicant had she been a male, given that the candidate is observed to be female. Such probability involves two worlds: an actual world where $A = a_0$ (the candidate is female) and a counterfactual world where for the same individual $A = a_1$ (the same candidate is male).

Counterfactual fairness [31] is a fine-grained variant of ETT conditioned on all features. That is, a prediction $\hat{Y}$ is counterfactually fair if under any assignment of values $X = x$,

$$\mathbb{P}[\hat{Y}_{A\leftarrow a_1} = \hat{y} \mid X = x, A = a_0] = \mathbb{P}[\hat{Y}_{A\leftarrow a_0} = \hat{y} \mid X = x, A = a_0]. \tag{16}$$

3 Tensions between fairness notions

It has been proved that there are incompatibilities between fairness notions. That is, it is not always possible for a predictor to satisfy specific fairness notions simultaneously [5, 10, 48, 35]. In presence of such incompatibilities, the predictor should make a trade-off to satisfy some notions on the expense of others or partially satisfy all of them. Incompatibility[5] results are well summarized by Mitchell et al. [35] as follows:

Statistical parity (independence) *versus* conditional use accuracy equality (sufficiency). Independence and sufficiency are incompatible, except when both groups (protected and non-protected) have equal base rates

[5] The term impossibility is commonly used as well.

or $\hat{Y}$ and Y are independent. Note, however, that $\hat{Y}$ and Y should not be independent since otherwise the predictor is completely useless. More formally,

$$\underset{\text{(independence)}}{\hat{Y} \perp A} \quad \text{AND} \quad \underset{\text{(strict sufficiency)}}{Y \perp A \mid \hat{Y}} \implies \underset{\text{(equal base rates)}}{Y \perp A} \quad \text{OR} \quad \underset{\text{(useless predictor)}}{\hat{Y} \perp Y}$$

It is important to mention here that this result does not hold for the relaxation of sufficiency, in particular, predictive parity. Hence, it is possible for the output of a predictor to satisfy statistical parity and predictive parity between two groups having different base rates.

Statistical parity (independence) *versus* equalized odds (separation). Similar to the previous result, independence and separation are mutually exclusive unless base rates are equal or the predictor $\hat{Y}$ is independent from the actual label Y [5]. As mentioned earlier, dependence between $\hat{Y}$ and Y is a weak assumption as any useful predictor should satisfy it. More formally,

$$\underset{\text{(independence)}}{\hat{Y} \perp A} \quad \text{AND} \quad \underset{\text{(strict separation)}}{\hat{Y} \perp A \mid Y} \implies \underset{\text{(equal base rates)}}{Y \perp A} \quad \text{OR} \quad \underset{\text{(useless predictor)}}{\hat{Y} \perp Y}$$

Considering a relaxation of equalized odds, that is, equal opportunity or predictive equality, breaks the incompatibility between independence and separation.

Equalized odds (separation) *versus* conditional use accuracy equality (sufficiency). Separation and sufficiency are mutually exclusive, except in the case where groups have equal base rates. More formally:

$$\underset{\text{(strict separation)}}{\hat{Y} \perp A \mid Y} \quad \text{AND} \quad \underset{\text{(strict sufficiency)}}{Y \perp A \mid \hat{Y}} \quad \Rightarrow \quad \underset{\text{(equal base rates)}}{Y \perp A}$$

Both separation and sufficiency have relaxations. Considering only one relaxation will only drop the incompatibility for extreme and degenerate cases. For example, predictive parity (relaxed version of sufficiency) is still incompatible with separation (equalized odds), except in the following three extreme cases [10]:

- both groups have equal base rates.
- both groups have $FPR = 0$ and $PPV = 1$.
- both groups have $FPR = 0$ and $FNR = 1$.

The incompatibility disappears completely when considering relaxed versions of both separation and sufficiency.

4 Tensions between fairness and privacy

The privacy property in the context of machine learning (ML) is typically formalized using differential privacy [18]. Differential privacy gives a strong guarantee that the learning algorithm will learn aggregate information about the population, but will not encode information about the individuals. Privacy and fairness of ML algorithms have been mainly studied separately. Recently, however, a number of studies focused on the relationship between fairness and privacy [14,38,2], that is, what is the consequence of guaranteeing fairness on the privacy of individuals? Also, to which extent the learning accuracy is impacted when fairness and privacy are simultaneously required? It turns out that there is a tension between privacy and fairness. In particular, it is impossible to satisfy exact fairness and differential privacy simultaneously while keeping a useful level of accuracy. Cummings et al. [14] provided a proof of a theorem stating that exact equal opportunity and differential privacy can simultaneously hold only for a constant/trivial classifier (a classifier that outputs always the same decision). However, the proof contains a flaw illustrated by Agarwal [2]. Agarwal, in turn, proved a stronger version of the theorem which holds on relaxed versions of fairness notions but could not fix Cummings et al.'s proof. This section describes a complete proof of the impossibility of satisfying simultaneously exact fairness and differential privacy while keeping a non-trivial accuracy.

For the sake of the proof, we use the same variable definitions as in Section 2. In addition, let $\mathcal{X}$ be the data universe consisting of all possible data elements $z = (x, a, y)$ where $x \in X$ are the element's features, $a \in A$ is the sensitive feature, and $y \in Y$ is the actual outcome (label). Let $h : \mathcal{X} \to \{0, 1\}$ be a binary classifier that tries to predict the true outcome y of a data element z.

The following definitions are needed for the proof.

Definition 1 (Trivial classifier) A classifier h is said to be trivial if it outputs the same outcome independently from the data element inputs:

$$\mathbb{P}[h(z) = \hat{y}] = \mathbb{P}[h(z') = \hat{y}] \qquad \forall z, z' \in \mathcal{X}, \quad \hat{y} \in \{0, 1\}$$

Definition 2 (Datasets adjacency) A dataset D can be defined in two ways each leading to a different definition of adjacency:

- a dataset is a finite set of samples $D = \{z_1, z_2, \ldots, z_n\}$ drawn from a distribution over $\mathcal{X}$. With this definition, datasets D and D' are adjacent if they differ in exactly one data element, that is, $z_i \neq z'_i$ for exactly one $i \in [n]$.
- a dataset is a distribution over $\mathcal{X}$. With this definition, D and D' are adjacent (ζ-close) if:

$$\frac{1}{2} \sum_{z \in \mathcal{X}} \mid D(z) - D'(z) \mid \leq \zeta,$$

where $D(z)$ is the probability of z under distribution D.

Definition 3 (Differential privacy) Let $\mathcal{D}$ be the set of all possible datasets and $\mathcal{R}$ the set of all possible trained classifiers. A learning algorithm $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ satisfies ϵ-differential privacy if for any two adjacent datasets D, $D' \in \mathcal{D}$, for any $\epsilon < \infty$, and for any subset of models $S \in \mathcal{R}$:

$$\mathbb{P}[\, \mathcal{M}(D) \in S \,] \;\; \leq \;\; e^{\epsilon}\, \mathbb{P}[\, \mathcal{M}(D') \in S \,]$$

Hence, to satisfy differential privacy, a learning algorithm should output similar classifiers with similar probabilities on any adjacent datasets.

Proposition 1 *Every trivial classifier is fair (equal opportunity) and differentially private.*

Proof We first prove that a trivial classifier satisfies always equal opportunity. Then we prove that it always satisfies differential privacy. Let h be a trivial classifier. Then,

$$\mathbb{P}[h(z) = 1] = \mathbb{P}[\hat{Y} = 1 | Y = y, A = a] \quad \forall z, y, a \tag{17}$$
$$= \mathbb{P}[\hat{Y} = 1 | Y = 1, A = 0] \tag{18}$$
$$= \mathbb{P}[\hat{Y} = 1 | Y = 1, A = 1] \tag{19}$$

Steps 18 and 19 correspond to equal opportunity (Equation 4).

For differential privacy, assume that the trivial classifier h outputs $\hat{Y} = 1$ with a constant probability $\rho \in]0, 1[$. Let D, $D' \in \mathcal{D}$ be two adjacent datasets. Then,

$$\forall z \in d \quad \mathbb{P}[\, h(z) = 1 \,] = \rho \tag{20}$$
$$\forall z' \in d' \quad \mathbb{P}[\, h(z') = 1 \,] = \rho \tag{21}$$

Hence, for any trivial classifier h

$$\mathbb{P}[\, \mathcal{M}(D) = h \,] = \mathbb{P}[\, \mathcal{M}(D') = h \,] \tag{22}$$

□

Proposition 2 *No learning algorithm $\mathcal{M}$ can simultaneously satisfy $\epsilon-$differential privacy and guarantee to generate a fair (equal opportunity) classifier which is non-trivial.*

To prove that Proposition 2 holds, it suffices to find a non-trivial classifier h which is fair on a dataset D and unfair on a neighboring dataset D'. This means that h can be generated by a model $\mathcal{M}$ on D but cannot be generated by the same model $\mathcal{M}$ on $D' \in \mathcal{D}$.

Proof [6] For any non-trivial classifier h, there exist two points a and b such that:

[6] The proof is inspired by Cummings et al. [14] and Agarwal [2] proofs.

– a and b are classified differently ($h(a) \neq h(b)$)[7]
– a and b belong to two different groups ($a = (x_1, 0, y_1)$ and $b = (x_2, 1, y_2)$)[8].

Consider datasets constructed over the following four elements:

$$z_1 = (x_1, 0, 1) \qquad z_2 = (x_1, 0, 0)$$
$$z_3 = (x_2, 1, 0) \qquad z_4 = (x_2, 1, 0)$$

Since h is non-trivial and depends only on the observable features (X and A), we have: $h(z_1) = h(z_2) = 0$ and $h(z_3) = h(z_4) = 1$. Let D a dataset over the above four points such that:

$$D(z_1) = \epsilon \qquad D(z_2) = \tfrac{1}{2} - \epsilon$$
$$D(z_3) = \epsilon \qquad D(z_4) = \tfrac{1}{2} - \epsilon$$

According to D, h is fair for group $A = 0$ (most of the points have label $Y = 0$ and are all classified $\hat{Y} = 0$) and for group for group $A = 1$ as well (most of the points have label $Y = 0$ and are all classified $\hat{Y} = 1$).

Consider now dataset D' on the same four points such that:

$$D'(z_1) = \tfrac{1}{2} - \epsilon \qquad D'z_2) = \epsilon$$
$$D'(z_3) = \tfrac{1}{2} - \epsilon \qquad D'(z_4) = \epsilon$$

According to D', h is (negatively) unfair to group $A = 0$ (most of the points have label $Y = 1$ but are all classified $\hat{Y} = 0$) and (positively) unfair to group $A = 1$ (most of the points have label $Y = 0$ but are all classified $\hat{Y} = 1$). It is important to mention finally that D and D' are not neighbors. However, according to Claim 2 in [2], if a learning algorithm is differentially private, then $\forall D, D' \in \mathcal{D}$, and for all classifiers h,

$$\mathbb{P}[\, \mathcal{M}(D) = h \,] > 0 \qquad \Longrightarrow \qquad \mathbb{P}[\, \mathcal{M}(D') = h \,] > 0 \tag{23}$$

which means that if h can be learned from dataset D, it can be also learned from dataset D'.

Hence, for any non-trivial classifier h which is fair on a dataset D, there always exist another dataset for which h is unfair. □

5 Tensions between process fairness and classification performance

This section focuses on explanation methods and how they are used to address the tensions between fairness and classification accuracy. It starts by discussing process fairness and its relation with fairness through unawareness [19]. It then recalls the main concepts underlying explanation methods. It finally shows how to use explanations to generate an ensemble classifier that allows to deal appropriately with the tension between fairness and classification accuracy.

[7] This is valid for any non-trivial classifier.

[8] If a and b belong to the same group, any point in the other group will be different from either a or b. So replace a or b with that point.

5.1 Process Fairness

Process fairness [24] can be described as a set of subjective fairness notions that are centered on the process that leads to outcomes. These notions are not focused on the fairness of the outcomes, instead they quantify the fraction of users that consider fair the use of a particular set of features. They are subjective as they depend on user judgments which may be obtained by subjective reasoning.

A natural approach to improve process fairness is to remove all sensitive (protected or salient) features before training classifiers. This simple approach connects process fairness to fairness through unawareness. However, in addition to the proxies problem mentioned in the beginning of Section 2, dropping out sensitive features may impact negatively classification performance [48]. Addressing the tension between these two constraints– classification performance and process fairness– requires to explore the set of classifiers that have suitable classification performance and at the same time low dependence on sensitive features.

A Rashomon set [20] is defined as a set of ML models that present similar performances in terms of error rate (the "good models") but that utilize features differently, e.g., they rely on class labels or certain features at different levels. Breiman [9] used "Rashomon effect" to denote a multiple functions with similar error rates but different descriptions. Recently, Coston et al. [13] adapted the notion of Rashomon set by integrating fairness metrics. We are interested in classifiers that belong to the set of "good" models, i.e. they have similar classification performance, but are less reliant on sensitive features. In order to quantify classifiers' reliance on sensitive features, we take advantage of explanation methods.

5.2 Explanation methods

Explanation methods differ mainly in the form of explanations or in the approach they use to generate them [25, 45]. The first group can be divided w.r.t. the type of explanations. For instance, Anchors provide rule-based explanations [40], while LIME [39], SHAP [33] and DeepLIFT [42] explain the outcome for a given instance by computing the contributions of every feature to the outcome. This group includes also methods based on group saliency maps [1], and counterfactual methods [47]. The second group can be arranged into two main sub-groups: (1) those that provide local explanations and (2) those that provide global explanations. Local explanation methods generate explanations for individual predictions, while global explanations give an understanding of the global behaviour of the model. Similarly, global explanation methods can be divided into methods based on a collection of local explanations [39], and representation based explanations [28].

In this survey, we are interested in explanation methods that are based on importance of features. Particularly, we focus on model agnostic explanation

methods that provide explanations in the form of feature importance such as LIME and SHAP.

5.3 Local explanations

Let f be the classifier, x be a target data instance, and $f(x)$ be the outcome we want to explain. In order to explain $f(x)$, LIME and SHAP generate data instances around x by applying perturbations. A mapping function $h_x(x')$ is responsible for converting x' from the interpretable space to the feature space. For instance, different data types require distinct mapping functions h_x. For tabular data, h_x treats discretized versions of numerical features, while for textual data, it deals with the presence/absence of words.

LIME and SHAP explanations take the form of surrogate models that are linear models (transparent by design). They learn a linear function g, i.e., $g(z') = w_g \cdot z'$, where w_g are the weights of the models which correspond to the importance of features. Now, let ξ be the explanation for $f(x)$, the function g optimize the following objective function:

$$\xi = \arg\min_{g \in \mathcal{G}} \{L(f, g, \pi_x) + \Omega(g)\}, \tag{24}$$

where $\Omega(g)$ measures the complexity of g (for instance, the arity of g) in order to insure interpretability of the linear model given by the explainer. L is the loss function defined by:

$$L(f, g, \pi_x) = \sum_{z, z' \in Z} [f(z) - g(z')]^2 \pi_x(z), \tag{25}$$

where z is the interpretable representation of x, and $\pi_x(z)$ defines the neighborhood of x that is considered to explain $f(x)$.

LIME and SHAP differs in the definition of the kernel π_x and also in the complexity function Ω used to produce explanations. In the following we recall each method and highlight their differences.

5.3.1 LIME

Local **I**nterpretable **M**odel Agostic **E**xplanations is a model-agnostic explanation method providing local explanations [39,23]. Explanations obtained from LIME take the form of surrogate linear models. LIME learns a linear model by approximating the prediction and feature values in order to mimic the behavior of ML model. To do so, LIME uses the following kernel π_x to define the neighborhood of x and considered to explain $f(x)$:

$$\pi_x(z) = exp(-d(x, z)^2/\sigma^2),$$

where d is a distance function between x and z, and σ is the kernel-width.

5.3.2 SHAP

SHapley **A**dditive ex**P**lanations [33] is also a local model-agnostic explanation method based on coalitional game theory. SHAP provides explanations in the form of a linear surrogate model that (unlike LIME) is defined on a simplified representation space (a "coalition" of simplified features), and whose coefficients correspond to the contributions of features. In the case of SHAP these coefficients coincide with Shapley values [41]. We focus on KernelSHAP [33] that is a variant of SHAP. KernelSHAP receives as input an instance x, the function f, and the number of coalitions m. It then learns a linear model g defined on a simplified subset of features ("coalition" that defines the representation space) by optimizing the loss function $L(f, g, \pi_x)$ with the kernel $\pi_x(z)$ defined as:

$$\pi_x(z) = \frac{K}{\binom{K}{|z|}|z|(K - |z|)},$$

where $|z|$ is the number of present features in the coalition z and K is the maximum coalition size.

KernelSHAP first samples coalitions of features and it then asks for prediction of each coalition. Before asking for predictions, KernelSHAP converts a coalition z from the representation space to the original space using $h_x(z)$. This produces a new dataset of coalitions along with predictions which is used by KernelSHAP to fit the linear model g.

Example. To illustrate, let us consider the example of the Adult dataset[9] where the goal is to predict if a person earns $\geq$50k dollars a year. The dataset contains more than 32000 instances; each instance is described by 14 features, e.g., "Age", "Education", and "Occupation". Figure 2 and 3 present LIME and SHAP explanations for a prediction using Logistic Regression classifier. In the case of SHAP explanation, the Shapley value for "Capital Gain = 2,174" is around -0.15 that indicates this feature contribute to move the prediction towards the negative class.

5.4 Assessing Fairness: From Local to Global Explanations

Local explanation methods only provide explanations for individual predictions. In order to assess fairness, we need to have a global understanding of the classifier. For instance, if classifier's outcomes depend on sensitive features, the classifier might be biased against a protected group. However, local explanations alone can not provide global understanding of the inner workings of classifiers. To overcome this issue, Ribeiro et al. [39] proposed the so-called Submodular-pick (SP). SP was originally proposed to work along with LIME explanations and it is called SP-LIME. The main idea on which relies SP-LIME is to sample a set of instances whose explanations are not redundant and that has a "high covering" in the following sense.

[9] http://archive.ics.uci.edu/ml/datasets/Adult

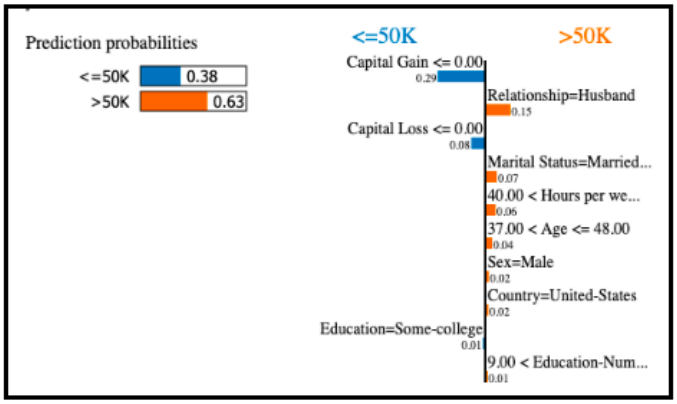


Fig. 2: LIME explanation of the prediction of an instance in the Adult dataset.

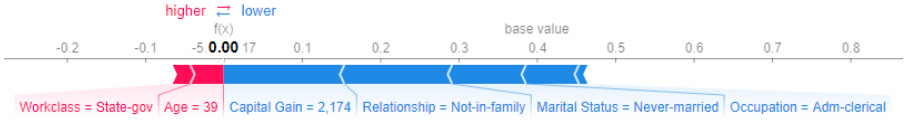


Fig. 3: SHAP explanation of the prediction of an instance in the Adult dataset.

Denote by $\mathcal{B}$ the desired number of explanations used to explain f globally, and let $\mathcal{V}$ be a set of selected instances, I an array of feature importance, and W an explanation matrix –columns represent features and rows represent instances– that contains the importance (contribution) of d' features to each instance. SP-LIME picks instances that are explaining thanks to:

$$Pick(W, I) = \arg \max_{\mathcal{V}, |\mathcal{V}| < \mathcal{B}} \sum_{j=1}^{d'} 1_{[i \in \mathcal{V}: \mathcal{W}_{ij} > 0]} I_j.$$

Once the number of desired explanations is attained, i.e., $\mathcal{V}$ is completed, then we aggregate the selected explanations. Essentially, for each feature in $\mathcal{V}$, we sum the contribution values of the respective feature that are present in all selected explanations. As a result, we obtain a single real number for each feature. This value represents the overall importance (contribution) of that feature to the classifier's outcomes. We call these feature contributions the global explanation.

Let $F^{(k)}$ be the list of the k most important features $a_1, a_2, ..., a_k$, where the absolute contribution of a_u is greater than a_v, $u < v$. If $F^{(k)}$ contains at least one sensitive feature $a_{j_1}, a_{j_2}, \ldots, a_{j_i}$ in $F^{(k)}$ with $i > 1$, then the classifier is deemed unfair.

5.5 Manipulating Explanations and Concealing Unfairness

Explanations can give insights about the inner workings of opaque classifiers. However, explanations based on feature importance can be manipulated. These manipulations can be done in such a way that biases are concealed even though explanations show that classifier's outcomes do not depend on sensitive features. This problem has been pointed out in recent papers [43, 17, 44].

Manipulating explanations can be done by adversarial attacks in the following way. An adversarial model is applied to train a biased classifier based on a particular fairness notion. An auditor uses explanations based on feature importance (LIME or SHAP) to assess (un)fairness.

LIME and SHAP generate data instances in the neighborhood of a data instance of interest in order to obtain local explanation. Generating data instances is done by perturbation and the distribution of generated data is not the same as the distribution of input data. Once adversarial models can differentiate between the two distributions, they can take advantage of it in the way they train/modify a classifier to be biased, in order to fool LIME and SHAP explanations.

5.6 Tackling Unfairness Through Unawareness: the case of FixOut

Algorithmic approaches that address unfairness issues are mainly divided in three groups based on the stage they apply fairness interventions [21]: *pre-processing*, *in-processing*, and *post-processing*. Pre-processing approaches modify the input to guarantee that the outcome is fair. In-processing techniques try to change the learning algorithm during the training process. Post-processing approaches modify algorithm's outputs to satisfy fairness constraints.

This section presents FixOut (FaIrness through eXplanations and feature dropOut), a framework that pushes further fairness through unawareness by combining in-processing and post-processing without compromising accuracy. FixOut removes sensitive features before training classifiers and modifies the input which characterizes the pre-processing phase. The framework produces a pool of classifiers whose outputs are combined thanks to an aggregation function. This function manipulates classifiers' outputs in order to enforce fairness, which characterizes the post-processing phase. More precisely, FixOut has two main components, namely: $\text{Exp}_{\text{Global}}$ and $\text{Ensemble}_{\text{Out}}$. $\text{Exp}_{\text{Global}}$ is responsible for assessing fairness of a pre-trained classifier. $\text{Ensemble}_{\text{Out}}$ is then applied if the pre-trained model is deemed unfair. It uses feature dropout, which manipulates the input with an ensemble approach to build a fair classifier.

FixOut receives a triple (M, D, F, E) of a pre-trained classifier M, a dataset D, a set of sensitive features F, and an explanation method E based on feature importance. It starts by applying the component $\text{Exp}_{\text{Global}}$ using E as the explanation method. For instance, it can employ either SHAP, LIME or any other measure of feature importance, and thus to evaluate the depen-

dence of M on sensitive features (see Figure 4). The output of $\text{Exp}_{\text{Global}}$ is a list $F^{(k)}$ of the k most important features $a_1, a_2, \ldots, a_k$. The framework applies the following rule to decide whether M is fair: if $F^{(k)}$ contains sensitive features $a_{j_1}, a_{j_2}, \ldots, a_{j_i}$ in F with $i > 1$, then M is deemed unfair and the FixOut's second component applies; otherwise, it is considered fair and no action is taken.

In the former case (i.e., M is considered unfair), FixOut employs *feature dropout* [7] and uses the i sensitive features $a_{j_1}, a_{j_2}, \ldots, a_{j_i} \in F^{(k)}$ to build a pool of $i+1$ classifiers in the following way:

- for each $1 \leq t \leq i$, FixOut trains a classifier M_t after removing a_{j_t} from D,
- and an additional classifier M_{i+1} trained after removing all sensitive features F from D.

This pool of classifiers is used to construct an ensemble classifier M_{final} (see Figure 4). As a post-processing approach, FixOut has considered three different aggregation functions for manipulating classifier's outputs in order to enforce fairness, namely: *simple*, *weighted* and *learned weighted* averages. Note that, FixOut uses only classifiers that provide probabilities.

Simple average. This is an immediate solution for aggregating classifiers' outputs. Here, all outputs have the same importance, even though some classifiers might be fairer than others. Given a data instance x and a class C, for an ensemble classifier M_{final} that uses simple averaging, the probability of x being in class C is computed as follows

$$P_{M_{final}}(x \in C) = \frac{1}{i+1} \sum_{t=1}^{i+1} P_{M_t}(x \in C), \tag{26}$$

where $P_{M_t}(x \in C)$ is the probability predicted by model M_t.

Weighted average. This function assigns different importance for classifiers' outputs. In order to to that, the contribution of sensitive features are taken into consideration. Let $c'_{j_t} \in [0, 1]$ be the normalized global feature contribution associated with a_{j_t}. We standardize feature contributions by $c'_{j_t} = \frac{c_{j_t} - min(F^{(k)})}{max(F^{(k)}) - min(F^{(k)})}$, where $min(F^{(k)})$ and $max(F^{(k)})$ are the lowest and the highest feature contribution among $F^{(k)}$, respectively. Now, let us define the weights w_t of M_t and the weight w_{i+1} of M_{i+1} as

$$w_t = \frac{c'_{j_t}}{1 + \sum_{u=1}^{i} c'_{j_u}},\ 1 \leq t \leq i, \quad \text{and} \quad w_{i+1} = \frac{1}{1 + \sum_{u=1}^{i} c'_{j_u}}.$$

The main idea behind using feature contribution in the weighted average is to ensure higher weights for classifiers trained without sensitive features whose contributions to M's outcomes are high. Also, the additional classifier M_{i+1}, the one that is trained without any sensitive feature, receives a higher weight.

For an ensemble classifier M_{final} that uses weighted averaging, the probability of x being in class C is computed as follows:

$$P_{M_{final}}(x \in C) = \sum_{t=1}^{i+1} w_t P_{M_t}(x \in C). \tag{27}$$

Learned weighted average. The third aggregation function also assigns different importance for classifiers' outputs. However, unlike the weighted average, weights are learned by a learning algorithm, e.g. Logistic Regression, instead of using directly contributions of sensitive features.

For each data instance, we ask the probabilities from each classifier in the pool. We then associate the list of probabilities obtained from all classifiers with the actual label. Thus, we have a dataset where each data instance is a list of probabilities with its label. This new dataset allows us to train a logistic regression classifier. After training, we use as weights the coefficients from the trained classifier. The following example illustrates how FixOut works.

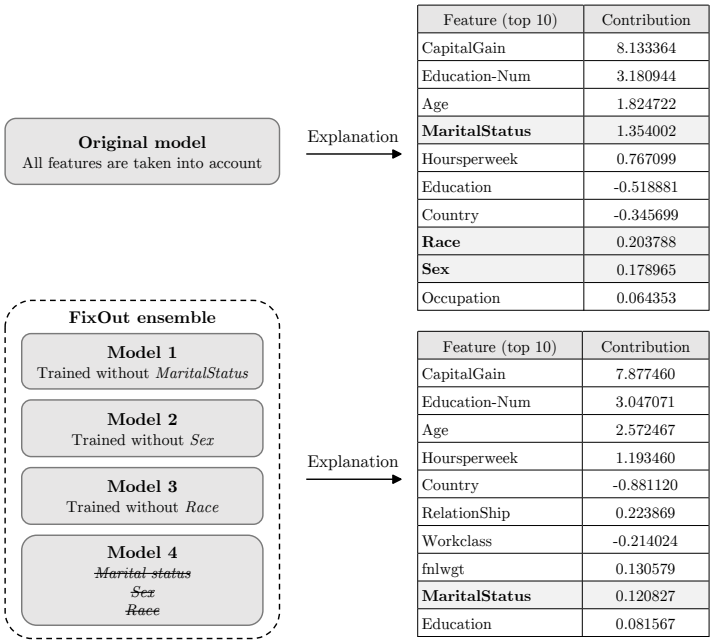
Example. We illustrate FixOut on the Adult dataset. The goal is to predict if an American citizen earns more than 50k dollars per year based on census information. In this dataset, the sensitive features are "MaritalStatus", "Race", and "Sex".

The global explanations and pool of classifiers obtained from the experiment are depicted in Figure 4. In the right side of the figure, we can see the ranking of features' contributions $F^{(k)}$, where $k = 10$ (also referred to here as the top-10 most important features), for both pre-trained (original model) and FixOut's ensemble classifiers. In the lower left part of the same figure, the pool of classifiers is shown. Note that, as we considered three features as sensitive features, FixOut trains four classifiers: three classifiers are trained without one sensitive feature (either "MaritalStatus", "Race", or "Sex") and a fourth one without any sensitive feature (all three sensitive features are removed before training).

Global explanations of the pre-trained classifier show that this classifier is dependent on sensitive features, i.e., all sensitive features have (absolute value of) contribution that place them in the top-10 most important features. On the other hand, global explanations of FixOut's ensemble show that the pool of classifiers obtained from feature dropout is less reliant of sensitive features. Note that, only "MaritalStatus" appears in the top-10 and this sensitive feature has lower contribution for the ensemble's outcomes than for the pre-trained classifier's outcomes.

5.7 FixOut for Textual Data

FixOut was extended to textual data and it was employed in the task of classifying tweets as hate speech or not [3]. The dataset used to evaluate this version contains tweets written in two language variant: African-American English and Standard-American English [15]. Classifiers trained on that dataset



Feature (top 10)	Contribution
CapitalGain	8.133364
Education-Num	3.180944
Age	1.824722
MaritalStatus	1.354002
Hoursperweek	0.767099
Education	-0.518881
Country	-0.345699
Race	0.203788
Sex	0.178965
Occupation	0.064353

Feature (top 10)	Contribution
CapitalGain	7.877460
Education-Num	3.047071
Age	2.572467
Hoursperweek	1.193460
Country	-0.881120
RelationShip	0.223869
Workclass	-0.214024
fnlwgt	0.130579
MaritalStatus	0.120827
Education	0.081567

Fig. 4: Impact of FixOut on global explanations of original model (pre-trained classifier) and FixOut's ensemble classifier. Example taken from an experiment on the Adult dataset using a bagging ensemble as pre-trained classifier and LIME.

were reported to be biased against tweets written in African-American English. For instance, some words that are considered offensive in Standard-American English are used in familiar interactions in African-American English, e.g. between close friends, and they do not indicate offensive discussions.

As language variant should not be a criteria for classifying a tweet as hate speech or not, FixOut was extended in order to reduce the dependence of classifiers on certain words. To do so, feature dropout was adapted to word dropout. However, once the number of words to be removed increase, FixOut becomes less effective. In order to overcome this issue, instead of ignoring a single word, words are grouped in order to do a "bag of word dropout", i.e., the classifier drops several words. The contribution of words using bag of words dropout (grouping words) is lower than word dropout (without grouping words).

6 Empirical analysis on benchmark datasets

To show how fairness notions are used to assess fairness and to illustrate some of the tensions described above, three benchmark datasets are used, namely, *communities and crimes, German credit,* and *Compas*. For each one

of them, the most common fairness notions are computed in two scenarios: baseline model (logistic regression including all the features in the dataset) and FixOut 's ensembles using logistic regression. This allows to highlight tensions between fairness notions and to show how feature dropping through process fairness produces an ensemble classifier with a good trade-off between fairness and classification accuracy.

6.1 Communities and crimes

The *communities and crimes* dataset[10] includes information relevant to per capita violent crime rates in several communities in the United States and the goal is to predict this crime rate. The dataset includes a total number of 123 numerical features and 1994 instances. 22 features have been dropped as they contain more than 80% missing values. The label *violent crime rate* has been transformed into a binary feature by thresholding[11] where 1 corresponds to high violent rate and 0 corresponds to low violent rate. To assess fairness, we consider two different settings depending on the sensitive feature at hand. First, the *communities racial makeup* is considered as the sensitive feature thus, two groups are created, namely: whites (communities with high rate of whites) and non-whites (communities with high rate of non-whites[12]). Second, the *communities rate of divorced female* is used as sensitive feature where we divide the samples into two sub-populations based on whether the rate of divorced females in a community is high (1) or low (0)[13].

Tables 1 and 2 show fairness assessment results for the *communities and crimes* dataset using the baseline model then FixOut. For both models, we applied the ten-fold cross-validation technique, using 90% of the data for training and the remaining 10% of the data for testing. Five fairness notions are applied, namely: statistical parity (SP), equal opportunity (EO), predictive equality (PE), predictive parity (PP) and calibration. Note that we binned the predicted scores in calibration in 10 bins and we calculated the bin-centers for each bin as shown in Table 2. The results show discrimination against communities with high rate of non-whites in the first setting and against communities with high rate of divorced females in the second setting for all fairness notions except for some of the calibration results corresponding to the bold-faced rates[14] presented in Table 2. Hence the only incompatibility exhibited by the experiment on *Communities and crime* is between sufficiency and independence in the case of few bins (bold-faced) in calibration results.

Process fairness empirical analysis focuses on the impact of feature dropout on classifiers' dependence on sensitive features. The results are shown in Table 3. Column "Contribution" contains the average value of feature contri-

[10] https://archive.ics.uci.edu/ml/datasets/communities+and+crime

[11] The mean value of the violent crime rate in the dataset is used as threshold.

[12] Blacks, Asians, Indians and Hispanics are grouped into a single group called non-whites

[13] The mean value of the divorced female rate in the dataset is used as threshold.

[14] We consider here a maximum difference of 0.01 as insignificant.

Table 1: Fairness assessment for the *communities and crimes* dataset.

Sensitive feature	Sub-population	Baseline SP	Baseline EO	Baseline PE	Baseline PP	FixOut SP	FixOut EO	FixOut PE	FixOut PP
Race	white	.14	.46	.05	.71	.14	.46	.05	.70
	non-white	.80	.86	.50	.85	.80	.89	.45	.86
Divorced female rate	high rate	.54	.72	.27	.82	.54	.72	.22	.82
	low rate	.08	.43	.00	.79	.08	.57	.03	.74

Table 2: Calibration obtained from experiments on the *communities and crimes* dataset.

Sensitive feature	Baseline *bin centers* .13	.21	.30	.38	.46	.54	.63	.71	.80	.88	FixOut *bin centers* .16	.26	.33	.40	.47	.54	.61	.66	.75	.82
Race	.08	**.20**	.29	.36	.45	.52	.60	.67	.75	.81	.08	**.20**	.29	.36	.44	.52	.60	.68	.75	.82
	.14	**.21**	.30	.41	.48	.56	.64	.71	.81	.90	.14	**.21**	.31	.42	.48	.56	.63	.72	.80	.91
Divorced female rate	.20	.24	.31	.39	**.46**	**.53**	**.60**	**.68**	.76	.87	.20	.24	.31	.39	**.45**	.52	**.60**	.68	.76	.87
	.01	.12	.28	.35	**.46**	**.54**	**.60**	**.67**	.72	.83	.06	.17	.28	.36	**.44**	.54	**.60**	.66	.73	.83

bution throughout the cross-validation protocol. Column "Ranking" presents the average position of features in the top k most important features; here, we adopted $k = 20$ for all experiments. We can observe that (absolute value of) contributions of both sensitive features decrease when we use FixOut, e.g., the absolute value of contribution of "Divorced female rate" decreases from 0.0199 (baseline) to 0.0080 (FixOut's ensemble). By analyzing the ranking, one notes that the position of both sensitive features decrease, i.e., the position in the list of most import features move down, which indicates that they become less important compared to other features (ranking positions increase). For instance, "Race" moved from 7.9 (baseline) to 15.5 position (FixOut's ensemble), i.e., it is closer to the end of the list. Classification accuracy for the FixOut ensemble classifier, however, remains exactly the same as the baseline case.

Table 3: Process fairness assessment for the *communities and crimes* dataset.

	Contribution		Ranking		Accuracy	
	Baseline	FixOut	Baseline	FixOut	Baseline	FixOut
Race	0.0092	0.0027	7.9	15.5	0.84	0.84
Divorced female rate	-0.0199	-0.0080	1.6	6.5		

6.2 German credit

The *German credit* dataset[15] is composed of the data of 1000 individuals applying for loans. Among 21 features in the dataset, 7 are numerical and 13 are categorical. Numerical and binary features are used directly as features in the classification and each categorical feature is transformed to a set of binary features, arriving at 27 features in total. This dataset is designed for binary classification to predict whether an individual will default on the loan (1) or not (0). We consider first, *gender* as sensitive feature where female applicants are compared to male applicants. Then, *age* is treated as protected feature where the population is divided into two groups based on whether they are above or below the mean age in the dataset (35.5 years-old).

Tables 4 and 5 show the results for assessing fairness notions for the *German credit* dataset. As for the communities and crimes dataset, two models are trained using 10-fold cross validation, namely, baseline and FixOut. Results for both models show that the applicants who are above the mean age are discriminated against compared to the applicants under the mean age based on SP, EO and PE. However, the results of PP show that the two sub-populations have almost the same predicted rate ($0.54 \approx 0.55$) regardless of the sensitive feature used (gender and age). That is, male and older applicants are privileged over female and younger applicants, respectively, when applying SP, EO and PE. However, there is parity when PP is used to assess fairness (around 0.55). Divergence between SP and PP is an example of the first incompatibility result in Section 3. Divergence between EO and PP is an example of the third incompatibility result in Section 3. Calibration results is inline with PP which confirms the first deviation.

Table 4: Fairness assessment for the German credit dataset.

Sensitive feature	Sub-population	Baseline				FixOut			
		SP	EO	PE	PP	SP	EO	PE	PP
Age	≥ 35.5	.11	.25	.03	**.55**	.15	.29	.10	.52
	< 35.5	.23	.45	.30	**.54**	.20	.38	.12	.59
Gender	female	.28	.75	.25	**.55**	.10	.23	.07	.52
	male	.14	.33	.16	**.55**	.21	.37	.13	.58

Table 6 shows the contribution of the sensitive feature on the classification output for the baseline as well as the FixOut models. Notice that the configuration is the same as the *communities and crime* case. Similarly to *communities and crime*, FixOut improves the contribution and ranking of "Age" compared to the baseline. However, FixOut only improved the ranking of "Gender" but not the contribution of this feature. Classification accuracy has slightly dropped from 0.71 in the baseline model to 0.69 in the FixOut model.

[15] https://archive-beta.ics.uci.edu/ml/datasets/statlog+german+credit+data

Table 5: Calibration obtained from experiments on the German credit dataset.

Sensitive feature	Baseline *bin centers*										FixOut *bin centers*									
	.23	.29	.35	.41	.47	.53	.60	.65	.71	.78	.16	.23	.31	.39	.46	.54	.62	.70	.77	.85
Age	.19	.22	.28	.35	.40	.45	.51	**.58**	**.63**	.68	**.19**	**.23**	**.31**	**.37**	**.43**	**.48**	**.54**	**.61**	**.67**	**.71**
	.23	.26	.32	.38	.43	.49	.54	**.58**	**.63**	.66	**.20**	**.24**	**.32**	**.38**	**.44**	**.49**	**.55**	**.61**	**.66**	**.70**
Gender	.21	.25	.31	.38	**.44**	.51	.57	**.62**	**.66**	**.71**	.20	.23	.29	.34	.40	.44	**.52**	**.56**	.64	.69
	.18	.23	.29	.36	**.43**	.47	.55	**.61**	**.66**	**.71**	.23	.27	.33	.39	.44	.48	**.52**	**.57**	.62	.66

Table 6: Process fairness assessment for the German credit dataset.

	Contribution		Ranking		Accuracy	
	Baseline	FixOut	Baseline	FixOut	Baseline	FixOut
Age	-0.0111	-0.0060	11.0	14.1	0.71	0.69
Gender	-0.0001	0.0020	15.0	17.6		

6.3 Compas

The *Compas* dataset contains information from Broward County, Florida, initially compiled by ProPublica [4] and the goal is to predict the two-year violent recidivism. That is, whether a convicted individual would commit a violent crime in the following two years (1) or not (0). Only black and white defendants who were assigned *Compas* risk scores within 30 days of their arrest are kept for analysis [4] leading to 5915 individuals in total. We consider *race* as sensitive feature in the first setting and *gender* in the second. Each categorical feature is transformed to a set of binary features leading to 11 features in total.

Similarly to the previous experiments, Tables 7 and 8 show the five fairness notions results for the baseline and the FixOut models. The tables show similar findings as those discussed in the *German credit* use case. That is, SP, EO and PE are not satisfied for both settings (blacks vs. whites and females vs. males) while the results of PP and calibration show closer results for both settings (0.69/0.65 for whites/black and 0.64/0.69 for females/males). This corroborates the debate that has arisen between Propublica and Northpointe[16] (*Compas* designers) where Propublica used EO and PE to prove that *Compas* privileges whites over blacks. At the other hand, the Northepointe's answer was that PP is a more suitable fairness notion to apply and they proved that *Compas* satisfies PP for blacks and whites [16].

For process fairness (Table 9), similarly to the previous benchmark datasets, the contribution of "Race" decreases when using FixOut's ensemble classifier. In the same way, the ranking of this feature increases from 7.1 in the case of the baseline model to 8.5 in the case of the FixOut ensemble classifier. Surprisingly, LIME explanations did not report "Gender" as a highly important feature for baseline's outcomes; this feature was already in the last position in the ranking (with no contribution). As a result, we do not see any decrease

[16] Now Equivant.

Table 7: Fairness assessment for the Compas dataset.

Sensitive feature	Sub-population	Baseline				FixOut			
		SP	EO	PE	PP	SP	EO	PE	PP
Race	white	.28	.42	.20	.65	.27	.42	.15	.66
	non-white	.56	.67	.36	.69	.53	.67	.34	.70
Gender	female	.23	.31	.13	.64	.46	.63	.18	.70
	male	.48	.66	.33	.69	.29	.70	.28	.62

Table 8: Calibration obtained from experiments on the Compas dataset.

	Baseline										FixOut									
Sensitive	*bin centers*										*bin centers*									
feature	.10	.19	.28	.37	.46	.54	.63	.72	.81	.90	.10	.19	.28	.37	.46	.54	.63	.72	.81	.90
Race	**.16**	**.21**	.27	**.37**	**.45**	**.53**	**.62**	**.71**	**.79**	**.86**	**.17**	.21	.27	**.36**	**.45**	**.53**	**.62**	.70	**.78**	**.86**
	.17	**.22**	.29	**.37**	**.46**	**.54**	**.63**	**.72**	**.79**	**.87**	**.17**	.23	.29	**.37**	**.46**	**.54**	**.63**	.72	**.79**	**.87**
Gender	**.16**	.20	**.27**	**.37**	**.44**	**.53**	**.62**	.70	**.79**	.85	**.17**	**.22**	**.28**	**.37**	**.45**	**.54**	**.63**	.72	**.79**	**.87**
	.17	.22	**.28**	**.37**	**.45**	**.54**	**.63**	.72	**.79**	.87	**.17**	**.21**	**.27**	**.37**	**.45**	**.53**	**.62**	.70	**.78**	**.86**

w.r.t feature contribution and ranking. Note finally that classification accuracy is almost the same (0.71 vs 0.70) for both models.

Table 9: Process fairness assessment for the Compas dataset.

	Contribution		Ranking		Accuracy	
	Baseline	FixOut	Baseline	FixOut	Baseline	FixOut
Race	-0.0017	-0.0003	7.7	8.5	0.71	0.70
Gender	0.0000	0.0000	10.0	10.0		

7 Conclusion

Implementing fairness is essential to guarantee that ML-based automated decision systems produce unbiased decisions and hence avoid unintentional discrimination against some sub-populations (typically minorities). This survey discusses two important issues related to implementing fairness.

First, there are several acceptable notions of fairness that can be impossible to satisfy simultaneously. This means that fairness practitioners have to choose among them. Second, implementing fairness can create tensions with other desirable properties of ML algorithms, in particular, privacy and classification accuracy. The survey also discusses process fairness which uses explanations and feature dropout to reduce dependence on sensitive features. This is a promising approach to improve fairness while keeping classification accuracy at an acceptable level.

Empirical results showed concrete examples of tensions between fairness notions in real datasets. For instance, for the *German credit* dataset, fairness

is satisfied according to predictive parity (sufficiency), but is not satisfied according to statistical parity (independence) and equalized odds (separation) which corroborates the incompatibility results of Section3. Empirical results showed also that using FixOut ensemble classifier (1) slightly reduced the disparity between sub-populations, (2) kept classification accuracy at similar level than the baseline case, and, most importantly, (3) significantly reduced the dependence of the output on the sensitive features.

This survey highlights the need to construct fair ML algorithms that address appropriately the different types of tensions. As future work, we also emphasize the need to automate the choice of sensitive/salient features in view of datasets and the decision task.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NeurIPS, pp. 9525–9536 (2018)
2. Agarwal, S.: Trade-Offs between Fairness and Privacy in Machine Learning. Master's thesis, University of Waterloo, Canada (2020)
3. Alves, G., Amblard, M., Bernier, F., Couceiro, M., Napoli, A.: Reducing unintended bias of ml models on tabular and textual data. The 8th IEEE International Conference on Data Science and Advanced Analytics (2021)
4. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. propublica. See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016)
5. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019). http://www.fairmlbook.org
6. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research (2018)
7. Bhargava, V., Couceiro, M., Napoli, A.: LimeOut: An Ensemble Approach To Improve Process Fairness. In: ECML PKDD Int. Workshop XKDD, pp. 475–491 (2020)
8. Bollen, K.A.: Structural equations with latent variables wiley. New York (1989)
9. Breiman, L.: Statistical modeling: The two cultures. Statistical science **16**(3), 199–231 (2001)
10. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
11. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018)
12. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 797–806 (2017)
13. Coston, A., Rambachan, A., Chouldechova, A.: Characterizing fairness over the set of good models under selective labels. arXiv preprint arXiv:2101.00352 (2021)
14. Cummings, R., Gupta, V., Kimpara, D., Morgenstern, J.: On the compatibility of privacy and fairness. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, pp. 309–315 (2019)
15. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17, pp. 512–515 (2017)
16. Dieterich, W., Mendoza, C., Brennan, T.: Compas risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc (2016)
17. Dimanov, B., Bhatt, U., Jamnik, M., Weller, A.: You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In: ECAI 2020, pp. 2473–2480. IOS Press (2020)

18. Dwork, C.: Differential privacy. In: International Colloquium on Automata, Languages, and Programming, pp. 1–12. Springer (2006)
19. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214–226 (2012)
20. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. JMLR **20**(177), 1–81 (2019)
21. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: ACM FAT, p. 329–338 (2019)
22. Gajane, P., Pechenizkiy, M.: On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184 (2017)
23. Garreau, D., von Luxburg, U.: Explaining the explainer: A first theoretical analysis of lime. In: AISTATS, pp. 1287–1296 (2020)
24. Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P., Weller, A.: The case for process fairness in learning: Feature selection for fair decision making. In: NIPS Symposium on Machine Learning and the Law, p. 2 (2016)
25. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
26. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29**, 3315–3323 (2016)
27. Kamiran, F., Zliobaite, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and information systems (Print) **35**(3), 613–644 (2013)
28. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: ICML, pp. 2668–2677 (2018)
29. Kim, M., Reingold, O., Rothblum, G.: Fairness through computationally-bounded awareness. In: Advances in Neural Information Processing Systems, pp. 4842–4852 (2018)
30. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent Trade-Offs in the Fair Determination of Risk Scores. In: C.H. Papadimitriou (ed.) 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), *Leibniz International Proceedings in Informatics (LIPIcs)*, vol. 67, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2017). DOI 10.4230/LIPIcs.ITCS.2017.43. URL http://drops.dagstuhl.de/opus/volltexte/2017/8156
31. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in Neural Information Processing Systems, pp. 4066–4076 (2017)
32. Lipton, Z., McAuley, J., Chouldechova, A.: Does mitigating ml's impact disparity require treatment disparity? In: Advances in Neural Information Processing Systems, pp. 8125–8135 (2018)
33. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: NIPS, pp. 4765–4774 (2017)
34. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)
35. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K.: Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. arXiv preprint arXiv:1811.07867 (2020)
36. Pearl, J.: Causality. Cambridge university press (2009)
37. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
38. Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., Miklau, G.: Fair decision making using privacy-protected data. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 189–199 (2020)
39. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the predictions of any classifier. In: ACM SIGKDD, pp. 1135–1144 (2016)

40. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: AAAI, pp. 1527–1535 (2018)
41. Shapley, L.S.: A value for n-person games. In: Contributions to the Theory of Games, pp. 307–317 (1953)
42. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. ICML pp. 3145–3153 (2017)
43. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180–186 (2020)
44. Slack, D., Hilgard, S., Singh, S., Lakkaraju, H.: Feature attributions and counterfactual explanations can be manipulated. CoRR **abs/2106.12563** (2021). URL https://arxiv.org/abs/2106.12563
45. Sokol, K., Flach, P.: One explanation does not fit all. KI-Künstliche Intelligenz **34**(2), 235–250 (2020)
46. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pp. 1–7. IEEE (2018)
47. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law & Technology **31**(2) (2018)
48. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th international conference on world wide web, pp. 1171–1180 (2017)
49. Zliobaite, I.: A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148 (2015)

A Notation Index

Table 10: Notation

V	set of attributes
A	sensitive attributes
X	remaining (non-sensitive) attributes
Y	actual outcome
$\hat{Y}$	outcome returned
S	score
R	resolving features
M	pre-trained classifier
D	dataset
F	list of features contributions
$F^{(k)}$	list of the k most important features
E	explanation method
x	data instance
$f(x)$	outcome of a classifier
g	linear (interpretable) model
z	interpretable representation of x
$h_x(z)$	transformation function
K	maximum coalition size
π	kernel (LIME,SHAP)
σ	kernel-width
Ω	measure of complexity
d	distance function
$\mathcal{B}$	desired number of explanations
$\mathcal{V}$	selected instances
W	explanation matrix
I	array of feature importance
w_t	weight of the t-th classifier
C	class (label)
a_i	i-th attribute (feature)
c_i	global feature contribution associated with a_i