HAL
open science

Survey on Fairness Notions and Related Tensions

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

To cite this version:

Guilherme Alves, Fabien Bernier, Miguel Couceiro, Karima Makhlouf, Catuscia Palamidessi, et al.. Survey on Fairness Notions and Related Tensions. EURO journal on decision processes, In press, 10.1016/j.ejdp.2023.100033 . hal-03484009v3

HAL Id: hal-03484009

https://hal.science/hal-03484009v3

Submitted on 19 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Survey on Fairness Notions and Related Tensions

Guilherme Alves[a,*], Fabien Bernier[a], Miguel Couceiro[a], Karima Makhlouf[b], Catuscia Palamidessi[b] and Sami Zhioua[b]

[a] *Université de Lorraine, CNRS, Inria N.G.E., LORIA, F-54000, Nancy, France*
[b] *Inria, École Polytechnique, IPP, Paris, 91120, France*

ARTICLE INFO

Keywords:
Fairness notion
Tension within fairness
Unfairness mitigation

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. However, ML-based decision systems are prone to bias, which results 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 and other desirable properties such as privacy and classification accuracy. This paper surveys the commonly used fairness notions and discusses the tensions among them with privacy and accuracy. Different methods to address the fairness-accuracy trade-off (classified into four approaches, namely, pre-processing, in-processing, post-processing, and hybrid) are reviewed. The survey is consolidated with experimental analysis carried out on fairness benchmark datasets to illustrate the relationship between fairness measures and accuracy in real-world scenarios.

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 (Dwork, Hardt, Pitassi, Reingold and Zemel, 2012), equalized odds (Hardt, Price and Srebro, 2016), etc.), individual-based notions (e.g., fairness through awareness (Dwork et al., 2012)), and causal-based notions (e.g., total effect (Pearl, 2009), counterfactual fairness (Kusner, Loftus, Russell and Silva, 2017)). As fairness is a social construct (Jacobs and Wallach, 2021) and an ethical concept (Tsamados, Aggarwal, Cowls, Morley, Roberts, Taddeo and Floridi, 2022), 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 (Barocas, Hardt and Narayanan, 2019; Mitchell, Potash, Barocas, D'Amour and Lum, 2020; Hardt et al., 2016; Berk, Heidari, Jabbari, Kearns and Roth, 2021; Kleinberg, Mullainathan and Raghavan, 2017). As a consequence, practitioners assessing and/or implementing fairness need to choose among them.

In addition to the tensions between notions ("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 (Dwork, 2006)) are at odds with each other (Cummings, Gupta, Kimpara and Morgenstern, 2019; Agarwal, 2020; Pujol, McKenna, Kuppam, Hay, Machanavajjhala and Miklau, 2020). That is, a learning algorithm that satisfies differential privacy is not guaranteed to generate a classifier that satisfies fairness, unless it has trivial accuracy.

Several methods have been introduced in the literature to reduce bias in the classification output. These methods fall into three categories, namely, pre-processing, in-processing, post-processing. This survey describes a representative set of methods from each category (Reweighting (Calders, Kamiran and Pechenizkiy, 2009) and FairBatch (Roh, Lee, Whang and Suh, 2020) for pre-processing, adversarial learning (Zhang, Lemoine and Mitchell, 2018), exponentiated gradient (Agarwal, Beygelzimer, Dudík, Langford and Wallach, 2018), and AdaFair (Iosifidis and Ntoutsi, 2019) for in-processing, and threshold optimizer (Hardt et al., 2016)

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.

*Corresponding author

guilherme.alves-da-silva@loria.fr (G. Alves); fabien.bernier@loria.fr (F. Bernier); miguel.couceiro@loria.fr (M. Couceiro); karima.makhlouf@inria.fr (K. Makhlouf); catuscia.palamidessi@inria.fr (C. Palamidessi); sami.zhioua@inria.fr (S. Zhioua)

ORCID(s): 0000-0002-5004-4429 (G. Alves); 0000-0002-1887-9725 (F. Bernier); 0000-0003-2316-7623 (M. Couceiro); 0000-0001-6318-0713 (K. Makhlouf); 0000-0003-4597-7002 (C. Palamidessi); 0000-0003-2029-175X (S. Zhioua)

and SBD (Fish, Kun and Lelkes, 2016) for post-processing). In addition, we consider a fourth category for methods that combine different fairness interventions (e.g. in-processing and post-processing). We describe a representative set of methods in this category, namely, LFR (Zemel, Wu, Swersky, Pitassi and Dwork, 2013), FAE (Iosifidis, Fetahu and Ntoutsi, 2019), and FixOut (Alves, Bhargava, Bernier, Couceiro and Napoli, 2022).

Reducing bias leads, typically, to a drop in classification accuracy. Therefore, all aforementioned methods can be considered as approaches to tackle the tension between fairness and classification accuracy.

Several surveys of the relatively recent field of ML fairness can be found in the literature (Berk et al., 2021; Mehrabi, Morstatter, Saxena, Lerman and Galstyan, 2021; Verma and Rubin, 2018; Mitchell et al., 2020; Zliobaite, 2015b; Gajane and Pechenizkiy, 2017). 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 classification accuracy. In addition, unlike other surveys, this paper features an experimental analysis based on fairness benchmark datasets. The aim is to show how different unfairness mitigation methods generate classifiers with different fairness but also different accuracy levels.

Section 2 briefly presents commonly used fairness notions spanning all categories (group, individual, and causal) along with their formal definitions. Related work is provided in Section 3. Section 4 describes the tensions and incompatibilities that exist among the various fairness notions. Section 5 shows that fairness and privacy properties are at odds with each other and present a complete formal proof of this incompatibility. Section 6 is a survey on unfairness mitigation methods which trackle the tension between fairness and classification accuracy. Section 7 shows how fairness notions can be applied on benchmark and real datasets, and illustrates some of the tensions described in the previous sections. Section 8 concludes and mentions potential directions for future work.

2. Related Work

With the increasing need for ethical concerns in decision-making systems that have serious implications on individuals and society, several survey papers have been proposed in the literature in the few recent years. In this section, we revisit these survey papers and highlight how our survey deviates from them.

Makhlouf et al. (Makhlouf, Zhioua and Palamidessi, 2021) compiled a survey about existing fairness notions and their main contribution consists of addressing the problem of the applicability of fairness notions to a given real-world scenario. To tackle this problem, the authors identified a set of fairness-related characteristics of the real-world scenario at hand. Then, they analyzed and studied the behavior of each fairness notion. The result of fitting these two elements together consists of a decision diagram that can be used as a roadmap to guide practitioners and policy makers to select the most suitable fairness notion in a specific setup. However, their survey does not consider conflicts between fairness and privacy. Moreover, our survey provides empirical results that illustrate the tensions between some fairness notions and how the problem of the trade-off between accuracy and fairness can be tackled through explanability and ensemble techniques.

Mehrabi et al. (Mehrabi et al., 2021) proposed a broader scope for their overview: in addition to concisely listing 10 definitions of fairness metrics, they discussed different sources of bias and different types of discrimination, they listed methods to mitigate discrimination categorized into pre-processing, in-processing, and post-processing, and they discussed potential directions for contributions in the field. However, they did not discuss any tensions that exist between fairness notions and tensions between fairness, accuracy and privacy which we discuss in depth in this survey.

Arrieta et al. (Arrieta, Díaz-Rodríguez, Del Ser, Bennetot, Tabik, Barbado, García, Gil-López, Molina, Benjamins et al., 2020) provided an exhaustive overview on explainability in ML. They proposed a novel definition of explainable ML that takes into account the audience for which the explanations are addressed. They also presented a detailed taxonomy of recent contributions related to the explainability of different ML models including explaining Deep Learning methods. Then, they discussed in depth the concept of responsible AI, which imposes the systematic adoption of several AI principles, namely: fairness, accountability and privacy. The survey does not discuss the tensions that might exist between those principles.

The survey of Mitchell et al. (Mitchell et al., 2020) includes an exhaustive list of group and individual fairness notions and outlines most of the impossibility results among them. They also discussed in detail a "catalogue" of choices and assumptions in the context of fairness to address the question of how social goals are formulated into a prediction (ML) problem. Again, their survey does not tackle the problem of tensions between fairness and other ethical considerations (privacy and explainability) in decision making systems as is studied in this paper.

Tsamados et al. (Tsamados et al., 2022) compiled an overview on the ethical problems in AI algorithms and the solutions that have been proposed in the literature. In particular, they provided a conceptual map of six ethical concerns raised by AI algorithms namely: inconclusive, inscrutable, misguided evidence, unfair outcomes, transformative effects, and traceability. The first three concerns refer to epistemic factors, the fourth and the fifth are normative factors and the fifth is relevant both for epistemic and normative factors. The epistemic factors are related to the relevance of the accuracy of the data while the informative factors refer to the ethical impact of AI systems. Although the survey explores a broad scope related to ethical concerns in AI, it remains at a conceptual level and does not address how these ethical concerns are implemented in practice and how they enter in conflict with each other in detail, which we explore in depth in this article.

Other works discussing the trade-off between fairess notions include the work by Kleinberg et al.(Kleinberg, Mullainathan and Raghavan, 2016) which discussed the suitability of specific fairness notions in a specific setup. In particular, they discussed the applicability of calibration and balance notions. The survey of Berk et al. (Berk et al., 2021) studied the trade-offs between different group fairness notions and between fairness and accuracy in a specific context namely: criminal justice risk assessments. They used simple examples based on the confusion matrix to highlight relationships between the fairness notions.

In another research direction, Friedler et al. (Friedler, Scheidegger and Venkatasubramanian, 2016) discussed tensions between group fairness and individual fairness. In particular, they defined two worldviews namely: WYSIWYG and WAE. The WYSIWYG (What you see is what you get) worldview assumes that the unobserved (construct) space and observed space are essentially the same while the WAE (we're all equal) worldview implies that there are no inherent differences between groups of individuals based on potential protected attributes.

A more recent survey by Fioretto et al. (Fioretto, Tran, Van Hentenryck and Zhu, 2022) provided the constraints under which differential privacy and fairness may be achieved simultaneously or having conflicting goals. In particular, they showed that individual fairness and differential privacy can be applied at the same time while group fairness and differential privacy are incompatible. They also examined how and why differential privacy may magnify bias and unfairness in two different settings namely: decision problems and learning tasks. Then, they reviewed existing mitigation measures for the fairness issues arising in these two settings.

3. 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 non-protected group, while $A = 1$ designates the 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.

To illustrate the various ML fairness notions, we use a simple job hiring scenario (Table 1). Each sample in the dataset has the following attributes: education level (numerical), job experience (numerical), age (numerical), marital status (categorical), gender (binary) and a label (binary). The sensitive attribute is the applicant's gender, that is, we are focusing on whether male and female applicants are treated equally. Table 1(b) presents the predicted decision (first column) and the predicted score value (second column) for each sample. The threshold value is set to 0.5.

Statistical parity (Dwork et al., 2012) (a.k.a., *demographic parity* (Kusner et al., 2017), *independence* (Barocas, Hardt and Narayanan, 2017), *equal acceptance rate* (Zliobaite, 2015a), *benchmarking* (Simoiu, Corbett-Davies, Goel et al., 2017), *group fairness* (Dwork et al., 2012)) 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 protected and unprotected groups should be equal. Statistical parity implies that:

$$\frac{TP + FP}{TP + FP + FN + TN}$$ [2]

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

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

In the example of Table 1, the calculated predicted acceptance rate of hiring male and female applicants is 0.57 (4 out of 7) and 0.4 (2 out of 5), respectively. Thus, statistical parity is not satisfied.

Conditional statistical parity (Corbett-Davies, Pierson, Feller, Goel and Huq, 2017) (a.k.a., *conditional discrimination-aware classification* (Kamiran, Zliobaite and Calders, 2013)) is a variant of statistical parity obtained by controlling on a set of resolving features (also called explanatory features in (Kamiran et al., 2013)). 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 while leading to a *legitimate* discrimination. Conditional statistical parity holds if:

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

In the example of Table 1, assuming job experience (denoted R) is a resolving variable, there is discrimination against females when $R = 2$ or $R = 3$ but no discrimination when $R = 8$ according to conditional statistical parity.

Equalized odds (Hardt et al., 2016) (a.k.a., *separation* (Barocas et al., 2017), *conditional procedure accuracy equality* (Berk et al., 2021), *disparate mistreatment* (Zafar, Valera, Gomez Rodriguez and Gummadi, 2017), *error rate balance* (Chouldechova, 2017)) considers both the predicted and the actual outcomes. The prediction is conditionally independent of the protected feature, given the actual outcome ($\hat{Y} \perp A \mid Y$). In other words, equalized odds requires

[1] A list of all terms used in this survey appears in Appendix A .

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

Table 1
A simple job hiring example. Y represents the data label indicating whether the applicant is hired (1) or rejected (0). $\hat{Y}$ is the prediction which is based on the score S. A threshold of 0.5 is used.

(a) Dataset

Gender	Education Level	Job Experience	Age	Marital Status	Y
Female 1	8	2	39	single	0
Female 2	8	2	26	married	1
Female 3	12	8	32	married	1
Female 4	11	3	35	single	0
Female 5	9	5	29	married	1
Male 1	11	3	34	single	1
Male 2	8	0	48	married	0
Male 3	7	3	43	single	1
Male 4	8	2	26	married	1
Male 5	8	2	41	single	0
Male 6	12	8	30	single	1
Male 7	10	2	28	married	1

(b) Prediction

Ŷ	S
1	0.5
0	0.1
1	0.5
0	0.2
0	0.3
1	0.8
0	0.1
0	0.1
1	0.5
1	0.5
1	0.8
0	0.3

that both sub-populations have the same true positive rate $TPR = \frac{TP}{TP+FN}$ and false positive rate $FPR = \frac{FP}{FP+TN}$:

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

Using the same example (Table 1), the TPR for male and female groups is 0.6 and 0.33, respectively, while the FPR is the same (0.5) for both groups. Consequently, the equalized odds does not hold, as female candidates are discriminated against.

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* (Hardt et al., 2016) (a.k.a., *false negative error rate balance* (Chouldechova, 2017)) 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 consider FP, equal opportunity is completely insensitive to the number of false positives. Equalized odds does not hold in the example of Table 1 as TPR is higher for males than females.

The second relaxed variant of equalized odds is called *predictive equality* (Corbett-Davies et al., 2017) (a.k.a., *false positive error rate balance* (Chouldechova, 2017)) 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. The example in Table 1 satisfies predictive parity.

Conditional use accuracy equality (Berk et al., 2021) (a.k.a., *sufficiency* (Barocas et al., 2017)) 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, conditional use accuracy equality conditions 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:

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

The calculated PPVs for male and female applicants in the hiring example (Table 1) are 0.75 and 0.5, respectively. NPVs for male and female applicants are both equal to 0.33. Overall the dataset in Table 1 does not satisfy conditional use accuracy equality.

Predictive parity (Chouldechova, 2017) (a.k.a., *outcome test* in (Simoiu et al., 2017)) 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. According to predictive parity, there is a discrimination against female candidates in the same example (Table 1).

Overall accuracy equality (Berk et al., 2021) is achieved when overall accuracy for both groups is the same. This implies that:

$$\frac{TP + TN}{TP + FN + FP + TN} \tag{8}$$

is equal for both groups:

$$\mathbb{P}[\hat{Y} = Y|A = 0] = \mathbb{P}[\hat{Y} = Y|A = 1] \tag{9}$$

Overall accuracy does not hold in Table 1 as $\mathbb{P}[\hat{Y} = Y]$ for female candidates is $\frac{2}{5}$ whereas for male candidates, it is $\frac{4}{7}$.

Treatment equality (Berk et al., 2021) is achieved when the ratio of FPs and FNs is the same for both protected and unprotected groups:

$$\frac{FN}{FP}_{\text{A=0}} = \frac{FN}{FP}_{\text{A=1}} \tag{10}$$

For example, in Table 1, the ratio $\frac{FN}{FP}$ for male is $\frac{2}{5}$ whereas for female, it is $\frac{2}{7}$.

Total fairness (Berk et al., 2021) 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 that is very difficult to hold in practice.

Balance (Kleinberg et al., 2017) uses the predicted probability 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:

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

In Table 1, the expected score values are 0.3 and 0.5 for hired males and hired females respectively. This also indicates discrimination against females since the latter need a higher score to get hired than males. *Balance of negative class* focuses instead on the negative class:

$$E[S \mid Y = 0, A = 0] = E[S \mid Y = 0, A = 1]. \tag{12}$$

The values in Table 1 are 0.35 and 0.3 for the non-hired males and females respectively. Hence, there is no balance for negative class, however, there is discrimination *in favor* of female.

Calibration (Chouldechova, 2017) (a.k.a. *test-fairness* (Chouldechova, 2017), *matching conditional frequencies* (Hardt et al., 2016)) holds if, for each predicted probability score $S = s$, individuals in all groups have the same probability to actually belong to the positive class:

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

Well-calibration (Kleinberg et al., 2017) is a stronger variant of calibration. It requires that (1) calibration is satisfied, (2) the score is interpreted as the probability of truly belonging to the positive class, and (3) for each score $S = s$, the probability of truly belonging to the positive class is equal to that particular score:

$$\mathbb{P}[Y = 1 \mid S = s, A = 0] = \mathbb{P}[Y = 1 \mid S = s, A = 1] = s \quad \forall\, s \in [0, 1]. \tag{14}$$

Table 2 shows the difference between calibration and well-calibration. The values inside the table represent $\mathbb{P}[Y = y \mid S = s, A = a]$.

Fairness through unawareness (a.k.a., *blindness, unawareness* (Mitchell et al., 2020), *anti-classification* (Corbett-Davies and Goel, 2018), *treatment parity* (Lipton, McAuley and Chouldechova, 2018).) is a simple and straightforward approach to address fairness problem where we ignore completely any sensitive feature while training the prediction system.

Fairness through awareness (Dwork et al., 2012) (a.k.a., *individual fairness* (Gajane and Pechenizkiy, 2017; Kusner et al., 2017)) 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 prediction outcomes. 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 (Kim, Reingold and Rothblum, 2018). A challenging aspect of this approach is the difficulty of determining an appropriate metric function to measure the similarity between two individuals. Typically, this requires careful human intervention from professionals with domain expertise (Kusner et al., 2017).

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 (Pearl, Glymour and Jewell,

Table 2
Calibration vs. well-calibration.

(a) Calibrated but not well-calibrated

s	0.4	0.7	0.8	0.85
Female	0.33	0.5	0.6	0.6
Male	0.33	0.5	0.6	0.6

(b) Calibrated and well-calibrated

s	0.4	0.7	0.8	0.85
Female	0.4	0.7	0.8	0.85
Male	0.4	0.7	0.8	0.85

2016). Causal relationships are expressed using structural equations (Bollen, 1989) 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.

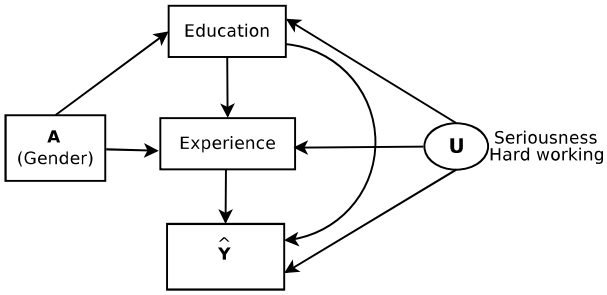

Figure 1: A possible causal graph for the hiring example.

Total effect (TE) (Pearl, 2009) is the causal version of statistical parity and is defined in terms of experimental probabilities as follows:

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

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 interventions. An intervention, denoted as $do(V = v)$, 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 as $\hat{Y}_{A=a}$ 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 that requires returning to the past and changing 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)* (Heckman, 1991) defined as:

$$ETT_{a_1,a_0}(\hat{y}) = \mathbb{P}[\hat{y}_{A\leftarrow a_1} \mid a_0] - \mathbb{P}[\hat{y} \mid a_0] \tag{16}$$

$\mathbb{P}[\hat{y}_{A\leftarrow a_1} \mid a_0]$ 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 has 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 the same individual $A = a_1$ (the same candidate is male).

Counterfactual fairness (Kusner et al., 2017) 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 $X = x$,

$$\begin{aligned}&\mathbb{P}[\hat{Y}_{A\leftarrow a_1} = \hat{y} \mid X = x, A = a_0] = \\ &\quad \mathbb{P}[\hat{Y}_{A\leftarrow a_0} = \hat{y} \mid X = x, A = a_0].\end{aligned} \tag{17}$$

It is important to mention that most of causal-based fairness notions require the knowledge of the causal graph. Causal graphs can be set manually by experts in the field, but very often generated using experiments (called also interventions). The process of identifying the causal graph is called causal discovery or structure learning. Binkyte et al. (Binkytė, Makhlouf, Pinzón, Zhioua and Palamidessi, 2023) show how different causal discovery approaches may result in different causal graphs and, most importantly, how even slight differences between causal graphs can have significant impact on fairness/discrimination conclusions.

Process fairness (Grgic-Hlaca, Zafar, Gummadi and Weller, 2016) (or *procedural fairness*) 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 at the beginning of Section 3, dropping out sensitive features may impact negatively classification performance (Zafar et al., 2017).

4. Tensions between fairness notions

It has been proved that there are incompatibilities between fairness notions. For instance, it is not always possible for a predictor to satisfy specific fairness notions simultaneously (Barocas et al., 2019; Chouldechova, 2017; Zafar et al., 2017; Mitchell et al., 2020). In the presence of such incompatibilities, the predictor should relax some fairness notions by partially satisfying all of them. Incompatibility[3] results are well summarized by Mitchell et al. (Mitchell et al., 2020) as follows. Before listing the tensions, it is important to summarize the relationships between fairness notions. In addition, for completeness, we define a new fairness notion, namely, *negative predictive parity*.

The following proposition formally states the relationship between equalized odds, equal opportunity, and predictive equality.

Proposition 1. *Satisfying equal opportunity and predictive equality is equivalent to satisfying equalized odds:*

$$Eq.\ 3 \Leftrightarrow Eq.\ 4 \wedge Eq.\ 5$$

Conditional use accuracy equality (Eq. 6) is "symmetric" to equalized odds (Eq. 3) with the only difference of switching Y and $\hat{Y}$. The same holds for equal opportunity (Eq. 4) and predictive parity (Eq. 7). However, there is no "symmetric" notion to predictive equality (Eq. 5). For completeness, we define such a notion and give it the name *negative predictive parity*.

Definition 1. *Negative predictive parity holds iff all subgroups have the same* $NPV = \frac{TN}{FN+TN}$*:*

$$P(Y = 1 \mid \hat{Y} = 0, A = 0) = P(Y = 1 \mid \hat{Y} = 0, A = 1) \quad (18)$$

Proposition 2. *Satisfying equalized odds or conditional use accuracy equality always leads to satisfying overall accuracy.*

$$Eq.\ 3 \vee Eq.\ 6 \Rightarrow Eq.\ (9)$$

The reverse, however, is not true. That is, a Machine Learning based Decision Making (MLDM) approach that satisfies overall accuracy does not necessarily satisfy equalized odds or conditional accuracy.

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 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,

$$\underset{\text{(independence)}}{\hat{Y} \perp A} \quad \wedge \quad \underset{\text{(strict sufficiency)}}{Y \perp A \mid \hat{Y}} \quad \Longrightarrow$$
$$\underset{\text{(equal base rates)}}{Y \perp A} \quad \vee \quad \underset{\text{(useless predictor)}}{\hat{Y} \perp Y}$$

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 of the actual label Y (Barocas et al., 2019). As mentioned earlier, dependence between $\hat{Y}$ and Y is a weak assumption as any useful predictor should satisfy it. More formally,

$$\underset{\text{(independence)}}{\hat{Y} \perp A} \quad \wedge \quad \underset{\text{(strict separation)}}{\hat{Y} \perp A \mid Y} \quad \Longrightarrow$$
$$\underset{\text{(equal base rates)}}{Y \perp A} \quad \vee \quad \underset{\text{(useless predictor)}}{\hat{Y} \perp Y}$$

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:

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

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 (Chouldechova, 2017):

- 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.

5. Tensions between fairness and privacy

Privacy in the context of machine learning (ML) is typically formalized using differential privacy (Dwork, 2006). Differential privacy gives a strong guarantee that the learning algorithm will learn aggregate information about the

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

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 (Cummings et al., 2019; Pujol et al., 2020; Agarwal, 2020). These studies attempt to answer two main questions: what is the consequence of guaranteeing fairness on the privacy of individuals? and 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. (Cummings et al., 2019) 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. On the other hand, the proof of Agrawal (Agarwal, 2020) does not contain flaws; although it holds onto a relaxed version of fairness, it does not address specifically equal opportunity. This section describes complete proof of the impossibility of satisfying simultaneously exact fairness (equal opportunity) and differential privacy while keeping a non-trivial accuracy. Hence, compared to the works of Cummings (Cummings et al., 2019) and Agarwal (Agarwal, 2020), our proof addresses the flaw of the former and shows explicitly how the proof holds specifically for equal opportunity, which the latter (Agarwal (Agarwal, 2020)) fails to address.

For the sake of the proof, we use the same variable definitions as in Section 3. 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 2 (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 3 (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 $D = \{z_1, z_2, \dots, z_n\}$ 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:*

$$\frac{1}{2} \sum_{z \in \mathcal{X}} \mid D(z) - D'(z) \mid \leq \zeta, \tag{19}$$

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

Equation 19 can be interpreted as a bounding of the Hamming distance between the two distributions D and D'. While the first definition of adjacency focuses on a specific data element z, the second definition (Equation 19) may involve several data elements $z_1, z_2, \dots \in \mathcal{X}$.

Definition 4 (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 $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ 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 $S \in \mathcal{R}$:*

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

$\mathbb{P}[\mathcal{M}(D) \in S]$ represents the probability that a learning algorithm outputs a classifier model in a specific subset of models S. Hence, to satisfy differential privacy, a learning algorithm should output similar classifiers with similar probabilities on any adjacent datasets.

Proposition 3. *Every trivial classifier is fair (equal opportunity) and differentially private.*

Proof 1. *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,*

$$\mathbb{P}[h(z) = 1] = \mathbb{P}[\hat{Y} = 1 | Y = y, A = a] \quad \forall z, y, a \tag{20}$$

$$= \mathbb{P}[\hat{Y} = 1 | Y = 1, A = 0] \tag{21}$$

$$= \mathbb{P}[\hat{Y} = 1 | Y = 1, A = 1] \tag{22}$$

Steps 21 and 22 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,

$$\forall z \in d \quad \mathbb{P}[\ h(z) = 1\ ] = \rho \tag{23}$$

$$\forall z' \in d' \quad \mathbb{P}[\ h(z') = 1\ ] = \rho \tag{24}$$

Hence, for any trivial classifier h

$$\mathbb{P}[\ \mathcal{M}(D) = h\ ] = \mathbb{P}[\ \mathcal{M}(D') = h\ ] \tag{25}$$

□

Proposition 4. *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 4 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 2. [4] *For any non-trivial classifier h, there exist two points a and b such that:*

- *a and b are classified differently* ($h(a) \neq h(b)$)[5]
- *a and b belong to two different groups* ($a = (x_1, 0, y_1)$ *and* $b = (x_2, 1, y_2)$)[6].

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:*

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

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* $A = 1$ *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:*

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

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 (Agarwal, 2020), if a learning algorithm is differentially private, then* $\forall D, D' \in \mathcal{D}$, *and for all classifiers h,*

$$\mathbb{P}[\ \mathcal{M}(D) = h\ ] > 0 \Longrightarrow \mathbb{P}[\ \mathcal{M}(D') = h\ ] > 0 \quad (26)$$

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 exists another dataset for which h is unfair. □

6. Tension between fairness and classification accuracy

This section focuses on the tension between fairness and classification accuracy, which is also known as the fairness-accuracy trade-off (Kleinberg et al., 2016). This tension naturally arises in many real-world scenarios, e.g., mortgage lending (Lee and Floridi, 2021). It is discussed in several papers (Bechavod and Ligett, 2017; Friedler, Scheidegger, Venkatasubramanian, Choudhary, Hamilton and Roth, 2019; Menon and Williamson, 2018; Zliobaite, 2015a) and it arises once we try to improve fairness in a ML pipeline by using a fairness processor.

Fairness processors are algorithmic approaches (also known as algorithmic interventions and fairness-enhancing interventions) that are conceived to optimize one or more fairness notions. These approaches are often arranged based on the stage they apply fairness interventions in a ML pipeline: *pre-processing*, *in-processing*, and *post-processing*. We give an overview of algorithmic interventions in this section. Particularly, we focus on methods whose implementations (source codes) are available and that cover all categories. A more complete list of fairness processors can be found at (Friedler et al., 2019; Pessach and Shmueli, 2022; Orphanou, Otterbacher, Kleanthous, Batsuren, Giunchiglia, Bogina, Tal, Hartman and Kuflik, 2022).

In this survey, we propose the inclusion of a fourth category named *hybrid-processing*, which comprises algorithmic approaches that combine different fairness interventions as a single method and, as a consequence, do not fit in any of the three traditional categories.

6.1. Pre-processing

Pre-processing approaches (pre-processors) modify the input in order to achieve fair outcomes. These processors can be applied to any model since they are model-agnostic. However, the fact they change the input before training may harm the explainability. In addition, pre-processors may increase the uncertainty of the classification process which impacts the level of accuracy (Pessach and Shmueli, 2022).

- *Reweighing* (Calders et al., 2009). This processor assigns different weights to data instances based on the distribution of a sensitive feature and the class label. The weights are used to guide a sampling procedure (with replacement) in order to obtain a (new) balanced dataset whose sensitive feature and class label are independent. For instance, data instances that obtained high weights will reappear more often in the balanced dataset. A classifier is then trained on the balanced dataset. As a consequence of the sampling procedure, classification errors on the high-weighted data instances are more expensive.

- *FairBatch* (Roh et al., 2020) This pre-processor is an extension of a batch selection algorithm that modifies the batch training in order to enforce model fairness (e.g., equal opportunity, equalized odds, demographic parity). More precisely, it measures fairness and adapts the size of the batch based on sensitive groups (which links this pre-processor with the reweighing approach).

[4] The proof is inspired by Cummings et al. (Cummings et al., 2019) and Agarwal (Agarwal, 2020) proofs.

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

[6] 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.

6.2. In-processing

In-processing techniques (in-processors) try to change the learning algorithm during the training process. Since they are an easy way to impose fairness constraints, these processors usually take into account the tension between fairness and classification performance. However, they can not always be applied to any model since they are usually model-specific.

- *Debiasing with adversarial learning* (Zhang et al., 2018). This in-processor trains two neural networks: (1) a predictor and (2) an adversary. Both networks have different objectives since the goal is, at the end of the training process, to attain the separation criterion. The goal of the predictor is to learn a function that predicts the class label (or the score in a regression problem), while the adversary takes as input the predicted label and its goal is to predict a sensitive feature. The predictor has weights W and loss function L_P, while the adversary has weights U and loss function L_A.

  The main idea behind this processor comes in the way the weights of both networks are updated. The weights U (of the adversary) are modified based only on the gradient $\nabla_U L_A$. Unlike the adversary's weights, the update of the predictor's weights, W, relies on two components: the first one is the gradient $\nabla_W L_P$ (that minimizes predictor's loss function), and the second one is $\text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A$ that avoids the predictor from helping or not trying to harm the adversary.

  This algorithmic intervention can be applied to classification and regression problems. Also, it can improve demographic parity, equalized odds, and equality of opportunity.

- *Exponentiated gradient* and *grid search reductions* (Agarwal et al., 2018; Agarwal, Dudik and Wu, 2019). These in-processors reduce an unfairness mitigation problem to a sequence of cost-sensitive classification problems. Here, fairness notions are re-written in a generic constraint format (*i.e.*, a vector of conditional moments), so that processors can support multiple fairness notions. More precisely, the statistical parity notion is a special case of the generic constraint that follows

  $$\mu_j(h) = \mathbb{E}[g_j(X, A, Y, h(X)) \mid \varepsilon_j],$$

  where g_j is a function that maps a data instance (protected and non-protected features) along with the predicted and actual outcomes into [0,1], h is a classifier, and ε_j is an event, which is independent of h, that relies on features (protected and non-protected) and Y. The idea is then to solve the following problem

  $$M\mu(h) \leq c, \tag{27}$$

  where M is a matrix $\mathcal{K} \times \mathcal{J}$, and c is a vector describing linear constraints. In order to empirically solve Eq. 27, $\mu(h)$ and c are replaced by $\hat{\mu}(h)$ and $\hat{c}$, respectively, in the form of a Lagrangian

  $$L(h, \lambda) = \mathbb{P}[h(X) = Y] + \lambda^T (M\hat{\mu}(h) - \hat{c}),$$

  where $\lambda \in \mathbb{R}_+^{|\mathcal{K}|}$ is a Lagrange multiplier. Now, one needs to find a saddle point that is the overlapped point between maximizing L and minimizing L. After a few iterations of updating λ, the optimal h is obtained as a result.

- *Prejudice remover regularizer* (Kamishima, Akaho, Asoh and Sakuma, 2012). This in-processor relies on the prejudice index (PI) that measures indirect prejudice. PI quantifies the mutual information between the class label and a sensitive feature, which indicates the degree of dependence on sensitive information.

  $$PI = \sum_{(Y,A)\in D} \mathbb{P}[Y, A] \ln \left(\frac{\mathbb{P}[Y, A]}{\mathbb{P}[Y]\mathbb{P}[A]} \right)$$

  PI is then included as a regularizer in the optimization function (see below), to take fairness into account. A parameter η is used to (reduce) enforce the importance of (un)fairness in the training process. The idea behind the penalty (along with the parameter η) is to reduce the dependency of the model on sensitive information and its fairness.

  $$\min_f L[f(x), Y] + \eta PI$$

  Prejudice remover can be applied to any discriminative probabilistic classifier. The original paper employed this in-processor in a logistic regression model.

- *AdaFair* (Iosifidis and Ntoutsi, 2019). It is an ensemble approach that trains a set of classifiers (weak learners) and then combines the output of these classifiers as a single outcome. More precisely, it is an extension of AdaBoost that takes into consideration Equalized Odds as a fairness notion to be optimized during training. In order to do that, it uses the notion of *cumulative fairness* which takes into account the fairness of a weak learner in the past iterations, instead of only considering the current iteration. In addition, AdaFair uses confidence scores in the re-weighting process in order to obtain fairer model outcomes, which differs from the traditional AdaBoost which relies only on classification error.

6.3. Post-processing

Post-processing approaches (post-processors) modify the algorithm's outputs to satisfy fairness constraints. They are usually model-agnostic and are an easier way to improve disparate impact. Also, post-processors are agnostic to the input data, which makes them easier to implement. However,

post-processors usually present low results compared to pre-processors (Kozodoi, Jacob and Lessmann, 2022; Pessach and Shmueli, 2022).

There exist two categories of post-processors: (1) transparent-box approaches that change the decision boundary of a model, and (2) opaque-box approaches that directly change the classification label.

- *Threshold Optimizer* (Hardt et al., 2016). It finds a threshold τ for each group (w.r.t. a sensitive feature) that minimizes a loss function L and, at the same time, takes into consideration the separation criterion (either equalized odds or equal opportunity). This post-processor can be employed on top of any model and it does not require information about non-sensitive features X, as it is based only on the joint distribution of the sensitive feature A, and the predicted and actual outcomes, $\hat{Y}$ and Y respectively. The optimization process searches in the receiver operating characteristic (ROC) curves between true and false positive rates of sub-populations. More precisely, for each binary-sensitive feature, this processor finds an intersection point between the ROC curves of the two sub-population, since the shape of the curves and the cost of misclassification are not necessarily the same for different groups. In other words, it minimizes the loss of classification while improving equalized odds (or equal opportunity) as described as follows

$$\min_{\tau} \mathbb{P}[S(X|A=a, Y=0) \leq \tau] \cdot L(\hat{Y}=1, Y=0) + (1 - \mathbb{P}[S(X|A=a, Y=1) > \tau]) \cdot L(\hat{Y}=0, Y=1).$$

- *Reject option classification* (Kamiran, Karim and Zhang, 2012). This post-processor requires that classifiers output probability scores. Data instances that obtain scores close to 0 or 1 indicate that a classifier has high confidence (low uncertainty) about predicting the class labels of those instances. Instead, the Reject option classification focuses on the other group of data instances. It relies on the idea that data instances that have high uncertainty can have their class labels switched in order to enforce fairness. Once the definition of high uncertainty (critical region) is established by the user, this processor changes the label of instances in the critical region to improve a certain fairness notion.

- *SBD* (Shifted Decision Boundary) (Fish et al., 2016). This post-processor is also inspired by the boosting mechanism. Unlike the traditional AdaBoost classifier that applies a majority voting to obtain the outcome, SBD uses a confidence score (and not classification error) for aggregating and obtaining the class label. Here, the statistical parity notion is incorporated into the confidence score, which is originally defined only by the distance of a data instance to the decision boundary. As a result, this strategy moves the boundary decision towards fairer outcomes (w.r.t. statistical parity). SBD can be employed on top of any model but it also allows the use of different fairness metrics.

6.4. Hybrid-processing

Hybrid-processing approaches (hybrid-processors) combine more than one algorithmic intervention in an ML pipeline. The idea is to use the advantages of a fairness processor to overcome the disadvantages of another processor.

- *LRF* (Learning Fair Representations) (Zemel et al., 2013) This fairness processor transforms (encodes) the data from an original space to a representation space in order to meet the following requirements: (1) to optimize statistical parity, (2) to lose the membership information about protected groups while keeping any other information from the original space, and (3) to map data instances from the representation space to Y such that the mapping has similar performance compared to an optimal classifier, i.e., to keep the highest possible accuracy. In order to do that, LRF takes into account these requirements by solving an objective function thanks to a in-processing approach. The goal is then to minimize the loss of a multi-objective and non-linear function.

- *FAE* (Fairness-Aware Ensemble) (Iosifidis et al., 2019) is a framework for fairness-aware classification that combines two fairness-enhancing interventions: pre-processing and post-processing. The first one tackles the problem of group imbalance and class imbalance by generating samples before the training phase (pre-processing intervention). It then trains multiple AdaBoost classifiers in order to obtain an ensemble model E. The second one moves the decision boundary of E towards fairer outcomes (post-processing intervention) based on the fairness notion Equal Opportunity. FAE sorts the misclassified instances (in descending order) in order to update the parameter θ that allows it to move the decision boundary.

- FIXOUT (FaIrness through eXplanations and feature dropOut) (Bhargava, Couceiro and Napoli, 2020; Alves, Amblard, Bernier, Couceiro and Napoli, 2021) is a framework that produces an ensemble model by combining pre-processing and post-processing interventions. At pre-processing, FIXOUT removes features based on the approach of fairness through unawareness. This intervention is guided by an explanation method (e.g., LIME (Ribeiro, Singh and Guestrin, 2016) and SHAP (Lundberg and Lee, 2017)) that produces a list $F^{(k)}$ of the k most important features $a_1, a_2, \ldots, a_k$ w.r.t a model M provided by the user. The framework then 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 fairness intervention 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* 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 then used to construct an ensemble classifier M_{final}. Note that, FIXOUT uses only classifiers that provide probabilities.

At the post-processing intervention, FIXOUT can use one of these aggregation functions in order to obtain a single outcome and to enforce fairness, namely: *simple*, *weighted*, and *learned weighted* averages.

In the simple average, 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

$$P_{M_{final}}(x \in C) = \frac{1}{i+1} \sum_{t=1}^{i+1} P_{M_t}(x \in C), \quad (28)$$

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

FIXOUT can improve fairness w.r.t. multiple sensitive features. However, it relies on explanations that can be vulnerable to adversarial attacks in some contexts.

Table 3 presents a summary of the fairness-enhancing interventions described here. It also indicates the links for the code artifacts (git repositories or the Python packages for fairness: AIF 360[11] or Fairlearn[12]).

6.5. Assumptions and expectations for the fairness-accuracy trade-off

Even though various fairness processors take into account both fairness and classification accuracy during the fairness intervention, there is still room for studying, characterizing, and defining this trade-off. On the one hand, distinct conclusions have been found about the impact on classification accuracy when fairness is enforced. For instance, one can say that improving fairness can compromise accuracy (Kleinberg et al., 2016), however, in specific contexts, it can actually increase accuracy (Pessach and Shmueli, 2021; Wick, Tristan et al., 2019).

On the other hand, other papers have been focused on characterizing or questioning the underlying assumptions made in previously published studies. For instance, (Zliobaite, 2015a) shows that in the evaluation of the fairness-accuracy trade-off, the acceptance rate must be taken into account since classification accuracy from distinct acceptance rates can not be comparable. More precisely, they rely on a notion of discrimination to assess fairness and show that better classification accuracy does not necessarily mean better classification if it comes from distinct acceptance rates. More recently, (Cooper, Abrams and NA, 2021) argue that researchers make assumptions that may lead to actually unfairness outcomes (or emergent unfairness). More precisely, three unsuitable assumptions are indicated: (1) fairness metrics are sufficient to assess fairness, (2) the lack of consideration of historical context, and (3) collecting more data on protected groups as an adequate solution.

7. 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 four scenarios: baseline model (logistic regression including all the features in the dataset), baseline model after Reweighing (pre-processor), baseline model along with Threshold Optimizer (post-processor), and hybrid-processor[13] (of pre-processing along with post-processing) using logistic regression. This allows us 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.

7.1. Communities and crimes

The *communities and crimes* dataset[14] 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[15] 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 a high rate of whites) and non-whites (communities with a high rate

[11] https://github.com/Trusted-AI/AIF360

[12] https://github.com/fairlearn/fairlearn

[7] https://github.com/iosifidisvasileios/AdaFair

[8] https://github.com/j2kun/fkl-SDM16

[9] https://github.com/iosifidisvasileios/Fairness-Aware-Ensemble-Framework

[10] https://gitlab.inria.fr/galvesda/fixout

[13] This hybrid-processor is equivalent to the unfairness mitigation step of FIXOUT.

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

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

Table 3
Summary of some fairness-enhancing interventions.

Type	Method	Fairness notion	Type of model	Artifact availability
Pre-proc.	Reweighing (Calders et al., 2009)	Statistical parity	Model agnostic	AIF 360
	FairBatch (Roh et al., 2020)	Equalized odds, Statistical parity	Model agnostic	-
	Debiasing with adversarial learning (Zhang et al., 2018)	Separation criterion	Gradient-based	AIF 360
In-proc.	Exponentiated Gradient (Agarwal et al., 2018, 2019)	Equalized odds, Statistical parity	Model agnostic	AIF 360, Fairlearn
	Prejudice Remover (Kamishima et al., 2012)	Normalized prejudice index	Logistic regression	AIF 360
	AdaFair (Iosifidis and Ntoutsi, 2019)	Equalized odds	AdaBoost	Git repository[7]
Post-proc.	Threshold Optimizer (Hardt et al., 2016)	Equalized odds	Any score based	AIF 360, Fairlearn
	Reject option classification (Kamiran et al., 2012)	Independence criterion	Model agnostic	AIF 360
	SBD (Fish et al., 2016)	Statistical parity	Any score based	Git repository[8]
Hybrid	LRF (Zemel et al., 2013)	Statistical parity	Logistic regression	AIF 360
	FAE (Iosifidis et al., 2019)	Equal opportunity	Bagging and boosting based	Git repository[9]
	FixOut (Bhargava et al., 2020; Alves et al., 2021)	Process fairness	Any score based	Git repository[10]

of non-whites[16]). Second, the *communities rate of divorced female* is used as a 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)[17].

Figures 2 and 3 show fairness assessment results for the *communities and crimes* dataset using the baseline model and then the hybrid-processor. 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), and predictive parity (PP). The results show discrimination against communities with a high rate of non-whites in the first setting and against communities with a high rate of divorced females in the second setting for all fairness notions. Fairness processors decrease the difference in some cases, in particular, EO was improved by the pre-processor and the post-processor, while the hybrid-processor reduces the discrimination w.r.t. to PP values in the first setting (Fig. 2). In the second setting (Fig. 3), the pre-processor and the post-processor have inverted the discrimination w.r.t. EO.

Process fairness empirical analysis focuses on the impact of feature dropout on classifiers' dependence on sensitive features. The results are shown in Table 4. Column "Contribution" contains the average value of feature contribution 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 (the absolute value of) contributions of both sensitive features decrease when we use the hybrid-processor (pre-processing and post-processing), e.g., the absolute value of the contribution of "Divorced female rate" decreases from 0.0199 (baseline) to 0.0080 (hybrid). By analyzing the ranking, one notes that the position of both sensitive features decreases, i.e., the position in the list of most import features moves 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 (hybrid), i.e., it is closer to the end of the list. Classification accuracy for the hybrid approach, however, remains exactly the same as the baseline

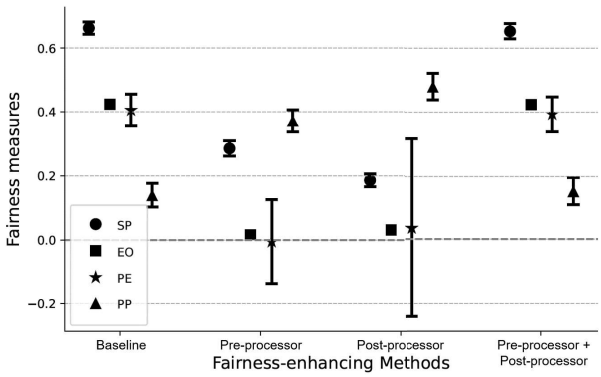

Figure 2: Fairness assessment for the communities and crimes dataset with race as a sensitive feature.

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

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

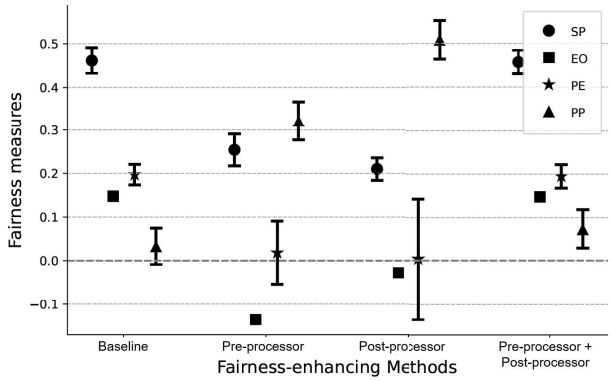

Figure 3: Fairness assessment for the communities and crimes dataset with divorced female rate as a sensitive feature.

case (see Figure 8) while the accuracy of the other two fairness processors dropped.

7.2. German credit

The *German credit* dataset[18] 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 a sensitive feature where female applicants are compared to male applicants. Then, *age* is treated as a 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).

Figures 4 and 5 show the results for assessing fairness notions for the *German credit* dataset. As for the communities and crimes dataset, four models are trained using 10-fold cross-validation, namely, baseline (logistic regression), baseline with a pre-processor, baseline with a post-processor, and the hybrid-processor (combination of pre-processor and post-processor). Overall, results for all methods show that the applicants who are above the mean age are discriminated against compared to the applicants under the mean age based on EO and PP 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 EO and PP. However, in the second setting (age as a sensitive feature, Fig. 5) there is parity when SP is used to assess fairness (SP close to 0) while the disparity of PP increases. Divergence between SP and PP is an example of the first incompatibility result in Section 4.

Table 5 shows the contribution of the sensitive feature on the classification output for the baseline as well as the hybrid approach. Notice that the configuration is the same as the

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

Table 4
Process fairness assessment for the *communities and crimes* dataset.

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

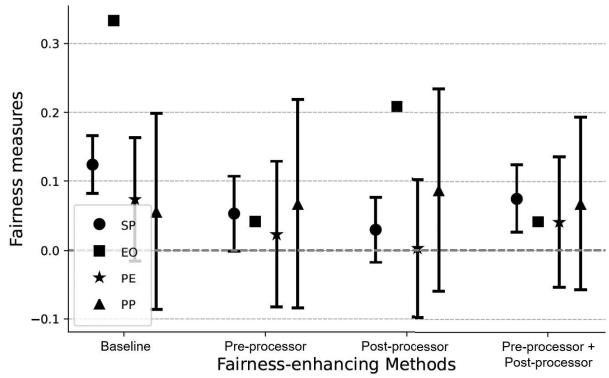

Figure 4: Fairness assessment for the German credit dataset with sex as a sensitive feature

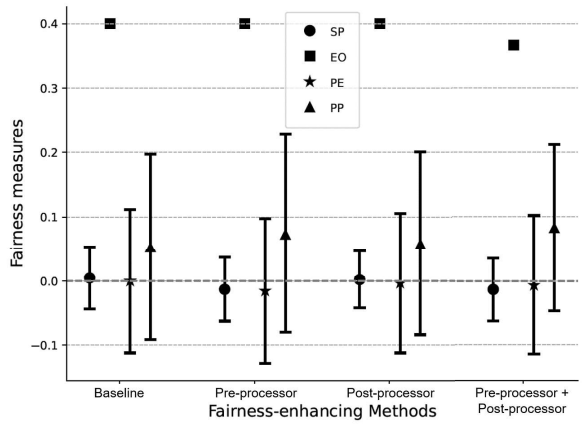

Figure 5: Fairness assessment for the German credit dataset with age as a sensitive feature

communities and crime case. Similarly to *communities and crime*, the hybrid-processor improves the contribution and ranking of "Age" compared to the baseline. However, the hybrid-processor 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 hybrid-processor.

7.3. Compas

The *Compas* dataset contains information from Broward County, Florida, initially compiled by ProPublica (Angwin, Larson, Mattu and Kirchner, 2016) and the goal is to predict the two-year violent recidivism. That is, whether a convicted

Table 5
Process fairness assessment for the German credit dataset.

	Contribution		Ranking	
	Baseline	Hybrid	Baseline	Hybrid
Age	-0.0111	-0.0060	11.0	14.1
Gender	-0.0001	0.0020	15.0	17.6

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 (Angwin et al., 2016) leading to 5915 individuals in total. We consider *race* as a sensitive feature in the first setting and *gender* in the second. Each categorical feature is transformed into a set of binary features leading to 11 features in total.

Similarly to the previous experiments, Figures 6 and 7 show the four fairness notions results for the baseline and the three fairness-enhancing methods (pre-processor, post-processor, and the hybrid combination of pre-processor and post-processor). The figures 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 we have at least one fairness notion closer to 0 (e.g. PP without any fairness intervention and PE after applying either the pre-processor or the post-processor). This corroborates the debate that has arisen between Propublica and Northpointe[19] (*Compas* designers) where Propublica used EO and PE to prove that *Compas* privileges whites over blacks. On 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 (Dieterich, Mendoza and Brennan, 2016).

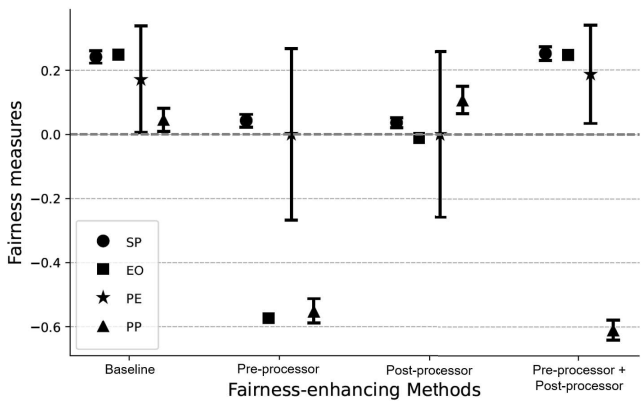

Figure 6: Fairness assessment for the Compas dataset with race as a sensitive feature.

For process fairness (Table 6), similarly to the previous benchmark datasets, the contribution of "Race" decreases when using the hybrid-processor. 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 hybrid-processor.

[19]Now Equivant.

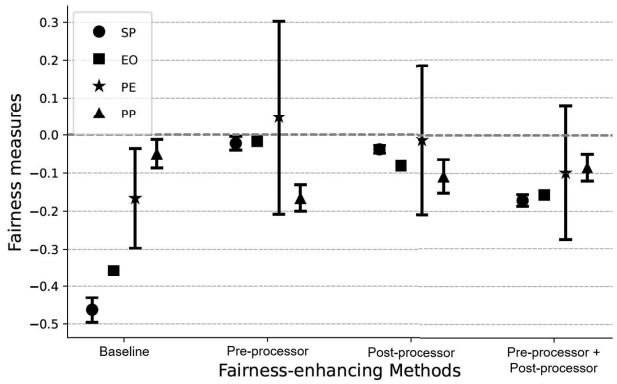

Figure 7: Fairness assessment for the Compas dataset with sex as a sensitive feature.

Surprisingly, LIME explanations did not report "Gender" as a highly important feature for baseline outcomes; this feature was already in the last position in the ranking (with no contribution). As a result, we do not see any decrease w.r.t feature contribution and ranking. Note finally that classification accuracy is almost the same for the hybrid-processor, the original model, and the pre-processor, but it decreased when the post-processor was applied (see Figure 8).

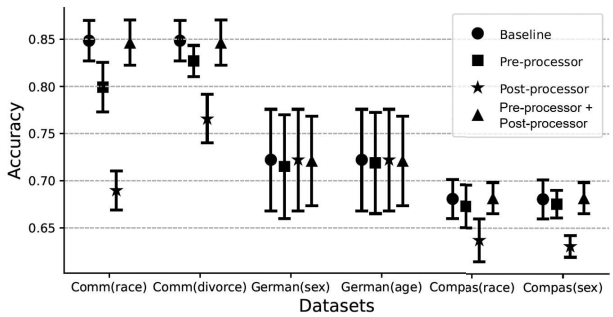

Figure 8: Accuracy of all datasets after applying the different fairness-enhancing methods.

8. 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 reasonable fairness requirements that can be satisfied simultaneously. This means that fairness practitioners have to choose among them. Second, implementing fairness can create tensions with other desirable properties of ML algorithms, such as privacy and classification accuracy. Empirical results showed that among the unfairness mitigation methods considered in this survey, pre-processing (reweighting) and post-processing (threshold

Table 6
Process fairness assessment for the Compas dataset.

	Contribution		Ranking	
	Baseline	Hybrid	Baseline	Hybrid
Race	-0.0017	-0.0003	7.7	8.5
Gender	0.0000	0.0000	10.0	10.0

optimizer) are the most efficient to mitigate bias. However, the hybrid approach produced the best accuracy among all unfairness mitigation methods. This survey highlights the need to construct fair ML algorithms that address appropriately the different types of tensions.

The most recent fairness notions in the literature are causal-based (Makhlouf, Zhioua and Palamidessi, 2020) and reflect the now widely accepted idea that using causality is necessary to appropriately address the problem of fairness. Hence, a promising future work is to study how the tensions described in this survey are reflected in causal-based fairness notions. For instance, enforcing causal-based fairness notions relax the tension with privacy and/or accuracy or amplify them. Besides, the most recent fairness notions, however, are causal-based and reflect the now widely accepted idea that using causality is necessary to appropriately address the problem of fairness.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H., 2018. A reductions approach to fair classification, in: International Conference on Machine Learning, PMLR. pp. 60–69.

Agarwal, A., Dudik, M., Wu, Z.S., 2019. Fair regression: Quantitative definitions and reduction-based algorithms, in: International Conference on Machine Learning, PMLR. pp. 120–129.

Agarwal, S., 2020. Trade-Offs between Fairness and Privacy in Machine Learning. Master's thesis. University of Waterloo. Canada.

Alves, G., Amblard, M., Bernier, F., Couceiro, M., Napoli, A., 2021. Reducing unintended bias of ML models on tabular and textual data. The 8th IEEE International Conference on Data Science and Advanced Analytics , 1–10.

Alves, G., Bhargava, V., Bernier, F., Couceiro, M., Napoli, A., 2022. Studying the impact of feature importance and weighted aggregation in tackling process fairness, in: Statistical Modeling and Applications on Real-Time Problems (to appear), CRC Press.

Angwin, J., Larson, J., Mattu, S., Kirchner, L., 2016. Machine bias. propublica. See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing .

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion 58, 82–115.

Barocas, S., Hardt, M., Narayanan, A., 2017. Fairness in machine learning. NIPS Tutorial .

Barocas, S., Hardt, M., Narayanan, A., 2019. Fairness and Machine Learning. fairmlbook.org. http://www.fairmlbook.org.

Bechavod, Y., Ligett, K., 2017. Learning fair classifiers: A regularization-inspired approach. CoRR abs/1707.00044. URL: http://arxiv.org/abs/1707.00044.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A., 2021. Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research 50, 3–44.

Bhargava, V., Couceiro, M., Napoli, A., 2020. LimeOut: An Ensemble Approach To Improve Process Fairness, in: ECML PKDD Int. Workshop XKDD, pp. 475–491.

Binkytė, R., Makhlouf, K., Pinzón, C., Zhioua, S., Palamidessi, C., 2023. Causal discovery for fairness, in: Proceedings of the Workshop on Algorithmic Fairness through the Lens of Causality and Privacy, PMLR. pp. 7–22. URL: https://proceedings.mlr.press/v214/binkyte23a.html.

Bollen, K.A., 1989. Structural equations with latent variables wiley. New York .

Calders, T., Kamiran, F., Pechenizkiy, M., 2009. Building classifiers with independency constraints, in: 2009 IEEE International Conference on Data Mining Workshops, IEEE. pp. 13–18.

Chouldechova, A., 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data 5, 153–163.

Cooper, A.F., Abrams, E., NA, N., 2021. Emergent unfairness in algorithmic fairness-accuracy trade-off research, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 46–54.

Corbett-Davies, S., Goel, S., 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 .

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A., 2017. 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.

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, pp. 309–315.

Dieterich, W., Mendoza, C., Brennan, T., 2016. Compas risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc .

Dwork, C., 2006. Differential privacy, in: International Colloquium on Automata, Languages, and Programming, Springer. pp. 1–12.

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, pp. 214–226.

Fioretto, F., Tran, C., Van Hentenryck, P., Zhu, K., 2022. Differential privacy and fairness in decisions and learning tasks: A survey. arXiv preprint arXiv:2202.08187 .

Fish, B., Kun, J., Lelkes, Á.D., 2016. A confidence-based approach for balancing fairness and accuracy, in: Proceedings of the 2016 SIAM international conference on data mining, SIAM. pp. 144–152.

Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., 2016. On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236 .

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: ACM FAT, p. 329–338.

Gajane, P., Pechenizkiy, M., 2017. On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184 .

Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P., Weller, A., 2016. The case for process fairness in learning: Feature selection for fair decision making, in: NIPS Symposium on Machine Learning and the Law, p. 2.

Hardt, M., Price, E., Srebro, N., 2016. Equality of opportunity in supervised learning. Advances in neural information processing systems 29, 3315–3323.

Heckman, J.J., 1991. Randomization and Social Policy Evaluation Revisited. Working Paper 107. National Bureau of Economic Research. URL: http://www.nber.org/papers/t0107, doi:10.3386/t0107.

Iosifidis, V., Fetahu, B., Ntoutsi, E., 2019. Fae: A fairness-aware ensemble framework, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE. pp. 1375–1380.

Iosifidis, V., Ntoutsi, E., 2019. Adafair: Cumulative fairness adaptive boosting, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 781–790.

Jacobs, A.Z., Wallach, H., 2021. Measurement and fairness, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 375–385.

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

Kamiran, F., Zliobaite, I., Calders, T., 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and information systems (Print) 35, 613–644.

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

Kim, M., Reingold, O., Rothblum, G., 2018. Fairness through computationally-bounded awareness, in: Advances in Neural Information Processing Systems, pp. 4842–4852.

Kleinberg, J., Mullainathan, S., Raghavan, M., 2016. Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807 .

Kleinberg, J., Mullainathan, S., Raghavan, M., 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores, in: Papadimitriou, C.H. (Ed.), 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany. pp. 43:1–43:23. URL: http://drops.dagstuhl.de/opus/volltexte/2017/8156, doi:10.4230/LIPIcs.ITCS.2017.43.

Kozodoi, N., Jacob, J., Lessmann, S., 2022. Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research 297, 1083–1094.

Kusner, M.J., Loftus, J., Russell, C., Silva, R., 2017. Counterfactual fairness, in: Advances in Neural Information Processing Systems, pp. 4066–4076.

Lee, M.S.A., Floridi, L., 2021. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. Minds and Machines 31, 165–191.

Lipton, Z., McAuley, J., Chouldechova, A., 2018. Does mitigating ml's impact disparity require treatment disparity?, in: Advances in Neural Information Processing Systems, pp. 8125–8135.

Lundberg, S.M., Lee, S., 2017. A unified approach to interpreting model predictions, in: NIPS, pp. 4765–4774.

Makhlouf, K., Zhioua, S., Palamidessi, C., 2020. Survey on causal-based machine learning fairness notions. arXiv preprint arXiv:2010.09553 .

Makhlouf, K., Zhioua, S., Palamidessi, C., 2021. On the applicability of machine learning fairness notions. SIGKDD Explor. 23, 14–23.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54, 1–35.

Menon, A.K., Williamson, R.C., 2018. The cost of fairness in binary classification, in: Conference on Fairness, Accountability and Transparency, PMLR. pp. 107–118.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K., 2020. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. arXiv preprint arXiv:1811.07867 .

Orphanou, K., Otterbacher, J., Kleanthous, S., Batsuren, K., Giunchiglia, F., Bogina, V., Tal, A.S., Hartman, A., Kuflik, T., 2022. Mitigating bias in algorithmic systems-a fish-eye view. ACM Computing Surveys (CSUR) .

Pearl, J., 2009. Causality. Cambridge university press.

Pearl, J., Glymour, M., Jewell, N.P., 2016. Causal inference in statistics: A primer. John Wiley & Sons.

Pessach, D., Shmueli, E., 2021. Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings. Expert Systems with Applications 185, 115667.

Pessach, D., Shmueli, E., 2022. A review on fairness in machine learning. ACM Computing Surveys (CSUR) 55, 1–44.

Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., Miklau, G., 2020. Fair decision making using privacy-protected data, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 189–199.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the predictions of any classifier, in: ACM SIGKDD, pp. 1135–1144.

Roh, Y., Lee, K., Whang, S.E., Suh, C., 2020. Fairbatch: Batch selection for model fairness, in: International Conference on Learning Representations.

Simoiu, C., Corbett-Davies, S., Goel, S., et al., 2017. The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics 11, 1193–1216.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., Floridi, L., 2022. The ethics of algorithms: key problems and solutions. AI & SOCIETY 37, 215–230.

Verma, S., Rubin, J., 2018. Fairness definitions explained, in: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), IEEE. pp. 1–7.

Wick, M., Tristan, J.B., et al., 2019. Unlocking fairness: a trade-off revisited. Advances in neural information processing systems 32.

Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P., 2017. 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.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C., 2013. Learning fair representations, in: International Conference on Machine Learning, pp. 325–333.

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

Zliobaite, I., 2015a. On the relation between accuracy and fairness in binary classification. arXiv preprint arXiv:1505.05723 .

Zliobaite, I., 2015b. A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148 .

A. Notation Index

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