Machine learning fairness notions: Bridging the gap with real-world applications

Karima Makhlouf, Sami Zhioua, Catuscia Palamidessi

▶ **To cite this version:**

Karima Makhlouf, Sami Zhioua, Catuscia Palamidessi. Machine learning fairness notions: Bridging the gap with real-world applications. Information Processing and Management, 2021, 58 (5), 10.1016/j.ipm.2021.102642 . hal-03624025

HAL Id: hal-03624025
https://hal.science/hal-03624025

Submitted on 13 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.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0306457321001321
Manuscript_fc01a0c56074f6184ab81f45e475e037

Machine Learning Fairness Notions: Bridging the Gap with Real-world Applications

Karima Makhlouf[a], Sami Zhioua[b], Catuscia Palamidessi[c,*]

[a] *Université du Québec à Montréal, Québec, Canada*
[b] *Higher Colleges of Technology, Dubai, UAE*
[c] *Inria, École Polytechnique, IPP, Paris, France*

Abstract

Fairness emerged as an important requirement to guarantee that Machine Learning (ML) predictive systems do not discriminate against specific individuals or entire sub-populations, in particular, minorities. Given the inherent subjectivity of viewing the concept of fairness, several notions of fairness have been introduced in the literature. This paper is a survey that illustrates the subtleties between fairness notions through a large number of examples and scenarios. In addition, unlike other surveys in the literature, it addresses the question of "which notion of fairness is most suited to a given real-world scenario and why?". Our attempt to answer this question consists in (1) identifying the set of fairness-related characteristics of the real-world scenario at hand, (2) analyzing the behavior of each fairness notion, and then (3) fitting these two elements to recommend the most suitable fairness notion in every specific setup. The results are summarized in a decision diagram that can be used by practitioners and policy makers to navigate the relatively large catalogue of ML fairness notions.

Keywords: Fairness, Machine learning, Discrimination, Survey , Systemization of Knowledge (SoK)

1. Introduction

Decisions in several domains are increasingly taken by "machines". These machines try to take the best decisions based on relevant historical data and using Machine Learning (ML) algorithms.

*Corresponding author.
Email addresses: karima.makhlouf@courrier.uqam.ca (Karima Makhlouf), szhioua@hct.ac.ae (Sami Zhioua), catuscia@lix.polytechnique.fr (Catuscia Palamidessi)

© 2021 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

Overall, ML-based decision-making (MLDM)[1] is beneficial as it allows to take into consideration orders of magnitude more factors than humans do and hence outputting decisions that are more informed and less subjective. However, in their quest to maximize efficiency, ML algorithms can systemize discrimination against a specific group of population, typically, minorities. As an example, consider the automated candidates selection system of St. George Hospital Medical School [1, 2]. The aim of the system was to help screening for the most promising candidates for medical studies. The automated system was built using records of manual screenings from previous years. During those manual screening years, applications with grammatical mistakes and misspellings were rejected by human evaluators as they indicate a poor level of English. As non-native English speakers are more likely to send applications with grammatical and misspelling mistakes than native English speakers do, the automated screening system built on that historical data ended up correlating race, birthplace, and address with a lower likelihood of acceptance. Later, while the overall English level of non-native speakers improved, the race and ethnicity bias persisted in the system to the extent that an excellent candidate may be rejected simply for her birthplace or address.

Given that MLDM can have a significant impact in the lives and safety of human beings, it is no surprise that social and political organization are becoming very concerned with the possible consequences of biased MLDM, and the related issue of lack of explanation and interpretability of ML-based decisions. The European Union has been quite active in this respect. Already in the General Data Protection Regulation (GDPR) there were directives concerning Automated Decision Making: for instance, Article 22 states that "The data subject shall have the right not to be subject to a decision based solely on automated processing." Other initiatives include the European Union's Ethics Guidelines for Trustworthy AI (April 2019), and OECD's Council Recommendation on Artificial Intelligence (May 2019).

In the scientific community, the issue of fairness in machine learning has become one of the most popular topics in recent years. The number of publications and conferences in this field has literally exploded, and a huge number of different notions of fairness have been proposed, leading sometimes to possible confusion. This paper, like other surveys in the literature (cf. Section 1), attempts to classify and systematize these notions. The characteristic of our work, however, consists in our point

[1] We focus on automated decision-making system supported by ML algorithms. In the rest of the paper we refer to such systems as MLDM.

of view, which is that *the very reason for having different fairness notions is how suitable each one of them is for specific real-world scenarios*. We feel that none of the existing surveys has addressed this aspect specifically. Discussion about the suitability (and sometimes the applicability) of the fairness notions is very limited and scattered through several papers [3, 4, 5, 6, 7, 8]. In this survey paper we show that each MLDM system can be different based on a set of criteria such as: whether the ground-truth exists, difference in base-rates between sub-groups, the cost of misclassification, the existence of a government regulation that needs to be enforced, etc. We then revisit exhaustively the list of fairness notions and discuss the suitability and applicability of each one of them based on the list of criteria.

Another set of results from the literature which is particularly related to the applicability problem we are addressing in this paper is the tensions that exist between some definitions of fairness. Several papers in the literature provide formal proofs of the impossibility to satisfy several fairness definitions simultaneously [3, 6, 8, 9, 10]. These results are revisited and summarized as they are related to the applicability of fairness notions.

The results of this survey are finally summarized in a decision diagram that hopefully can help researchers, practitioners, and policy makers to identify the subtleties of the MLDM system at hand and to choose the most appropriate fairness notion to use, or at least rule out notions that can lead to wrong fairness/discrimination result.

The paper is organized as follows. Section 3 lists notable real-world MLDMs where fairness is critical. Section 4 identifies a set of fairness-related characteristics of MLDMs that will be used in the subsequent sections to recommend and/or discourage the use of fairness notions. Fairness notions are listed and described in the longest section of the survey, Section 5. Section 6 discusses relaxations of the strict definitions of fairness notions. Section 7 describes classification and tensions that exist between some fairness notions. The decision diagram is provided and discussed in Section 8.

2. Related Work and Scope

With the increasing fairness concerns in the field of automated decision making and machine learning, several survey papers have been published in the literature in the few previous years. This section revisits these survey papers and highlights how this proposed survey deviates from them.

In 2015, Zliobaite compiled a survey about fairness notions that have been introduced previously [11]. He classified fairness notions into four categories, namely, statistical tests, absolute

measures, conditional measures, and structural measures. Statistical tests indicate only the presence or absence of discrimination. Absolute and conditional measures quantify the extent of discrimination with the difference that conditional measures consider legitimate explanations for the discrimination. These three categories correspond to the group fairness notions in this survey. Structural measures correspond to individual fairness notions[2]. Most of the fairness notions listed by Zliobaite are variants of the group fairness notions in this survey. For instance, difference of means test (Section 4.1.2 in [11]) is a variant of balance for positive class (Section 5.7 in this paper). Although, he dedicated one category for individual notions (structural measures), Zliobaite did not mention important notions, in particular fairness through awareness. Regarding the applicability of notions, the only criterion considered was the type of variables (e.g. binary, categorical, numerical, etc.).

The survey of Berk et al. [12] listed only group fairness notions that are defined using the confusion matrix. Similar to this survey, they used simple examples based on the confusion matrix to highlight relationships between the fairness notions. The applicability aspect has not been addressed as the paper focused only on criminal risk assessment use case.

The survey of Verma and Rubin [13] described a list of fairness notions similar to the list in this survey. To illustrate how each notion can be computed in real scenarios, they used a loan granting real use case (German credit dataset [14]). Rather than using a benchmark dataset, this survey uses a smaller and fictitious use case (job hiring) which allows to illustrate better the subtle differences between the fairness notions. For instance, counterfactual fairness is more intuitively described using a small job hiring example than the loan granting benchmark dataset. Verma and Rubin did not address the applicability aspect in their survey.

Gajane and Pechenizkiy [4] focused on formalizing only notable fairness notions (e.g. statistical parity, equality of opportunity, individual fairness, etc.) and discussed their implications on distributive justice from the social sciences literature. In addition, they described two additional fairness notions that are studied extensively in the social sciences literature, namely, equality of resources and equality of capability. These notions, however, do not come with a mathematical formalization. This survey is more exhaustive as it analyzes a much larger number of fairness notions. However, being focused on the implication on distributive justice, Gajane and Pechenizkiy's survey addresses the suitability of the discussed fairness notions in real world domains.

[2]Zliobaite does not use group vs individual notions, but indirect and direct discrimination.

Mehrabi et al. [15] considered a more general scope for their survey: in addition to briefly listing 10 definitions of fairness notions (Section 4.2), they surveyed different sources of bias and different types of discrimination, they listed methods to implement fairness categorized into pre-processing, in-processing, and post-processing, and they discussed potential directions for contributions in the field. This survey is more focused on fairness notions which are described in more depth.

A more recent survey by Mitchell et al. [3] presents an exhaustive list of fairness notions in both categories (group and individual) and summarizes most of the incompatibility results in the literature. Although Mitchell et al. discuss a "catalogue" of choices and assumptions in the context of fairness, the aim of these choices and assumptions is different from the criteria defined in this survey (Section 4). The assumptions and choices discussed in Section 2 in [3] address the question of how social goals are abstracted and formulated into a prediction (ML) problem. In particular, how the choice of the prediction goal, the choice of the population, and the choice of the decision space can have an impact on the degree of fairness of the prediction. Whereas the choices and criteria discussed in this survey (Section 4) are used to help identify the most suitable fairness notion to apply in a given scenario.

Other surveys include the one by Friedler et al. [16] which considered only group fairness notions and focused on surveying algorithms to implement fairness.

Overall most of existing review papers do not address all flavors of fairness notions in the same survey. In particular, most of them focus on statistical and group fairness notions. Causality based fairness notions, however, are not covered in several surveys while it is the most reliable category of notions in the disparate treatment legal framework. However, the main contribution of this survey is the focus on the applicability of fairness notions and the identification of fairness-related criteria to help select the most suitable notion to use given a scenario at hand. Brief discussions about the suitability of specific fairness notions can be found in few papers. For instance, Zafar et al. [5] mentioned some application scenarios for statistical parity and equalized odds. Kleinberg et al.[6] discussed the applicability of calibration and balance notions. Through a discussion about the cost of unfair decision on society, Corbett-Davies et al.[7] analyzed the impact of using statistical parity, predictive equality, and conditional statistical parity on public safety (criminal risk assessment). Gajane and Pechenizkiy [4] discuss the suitability of notable fairness notions (statistical parity, individual fairness, etc.) from the distributive justice point of view. Unlike the scattered discussions about the applicability of fairness notions found in the literature, this survey provides a complete

reference to systemize the selection procedure of fairness notions. A short version of this paper was presented in BIAS 2020 workshop at ECMLPKDD 2020 [17].

Fairness in machine learning can be categorized according to two dimensions, namely, the task and the type of learning. For the first dimension, there are two tasks in fairness-aware machine learning: discrimination discovery (or assessment) and discrimination removal (or prevention). Discrimination discovery task focuses on assessing and measuring bias in datasets or in predictions made by the MLDM. Discrimination removal focuses on preventing discrimination by manipulating datasets (pre-processing), adjusting the MLDM (in-processing) or modifying predictions (post-processing). For the second dimension, fairness can be investigated for different learning types including fairness in classification, fairness in regression [18, 19], fairness in ranking [20], fairness in reinforcement learning [21], etc. This survey focuses on the task of discrimination discovery (assessing fairness) in "pure prediction" [22] classification problems with a single decision making task (not sequential) and where decisions do not impact outcomes [23].

3. Real-world scenarios with critical fairness requirements

As the paper is focusing on the applicability of fairness notions, we provide here a list of notable real-world MLDMs where fairness is critical. In each of these scenarios, failure to address the fairness requirement will lead to unacceptable biased decisions against individuals and/or sub-populations. These scenarios will be used to provide concrete examples of situations where certain fairness notions are more suitable than others.

Job hiring: MLDMs in hiring are increasingly used by employers to automatically screen candidates for job openings[3]. Commercial candidate screening MLDMs include XING[4], Evolv [26], Entelo, Xor, EngageTalent, GoHire and SyRI [5]. Typically, the input data used by the MLDM include: affiliation, education level, job experience, IQ score, age, gender, marital status, address, etc.

[3] In 2014, the automated job screening systems market was estimated at $500 million annual business and was growing at a rate of 10 to 15% per year [24]

[4] A job platform similar to LinkedIn. It was found that this platform ranked less qualified male candidates higher than more qualified female candidates [25].

[5] System Riscico Indicatie, or SyRI for short, is a risk profiling system being deployed in the Netherlands by the Department of Social Affairs and Employment with the intention of identifying individuals who are at a high risk of committing fraud in relation to employment and other matters like social security and taxes. Its use raised a lot of controversy, and its case was brought to the Court of the Hague, that concluded on the 5th of February 2020 that the Government's use of SyRI violates the European Convention on Human Rights. To a very large extent, the Court's judgment was based on the lack of transparency in the algorithm at the heart of the system.

The MLDM outputs a decision and/or a score indicating how suitable/promising the application is for the job opening. A biased MLDM leads to rejecting a candidate because of a trait that she cannot control (gender, race, sexual orientation, etc.). Such unfairness causes a prejudice on the candidate but also can be damaging for the employer as excellent candidates might be missed.

Granting loans: Since decades, statistical and MLDM systems are used to assess loan applications and determine which of them are approved and with which repayment plan and annual percentage rate (APR). The assessment proceeds by predicting the risk that the applicant will default on her repayment plan. Loan Granting MLDMs currently in use include: FICO, Equifax, Lenddo, Experian, TransUnion, etc. The common input data used for loan granting include: credit history, purpose of the loan, loan amount requested, employment status, income, marital status, gender, age, address, housing status and credit score. An unfair loan granting MLDM will either deny a deserving applicant a requested loan, or give her an exorbitant APR, which on the long run will create a vicious cycle as the candidate will be very likely to default on her payments.

College admission: Given the large number of admission applications, several colleges are now resorting to MLDMs to reduce processing time and cut costs[6]. Existing college admission MLDMs include GRADE [27], IBM Watson[7], Kira Talent[8] . Typically, the candidates' features used include: the institutions previously attended, SAT scores, extra-curricular activities, GPAs, test scores, interview score, etc. The predicted outcome can be a simple decision (admit/reject) or a score indicating the candidate's potential performance in the requested field of study [10]. Unfair college admission MLDMs may discriminate against a certain ethnic group (e.g. African-American [28]) which could lead, in the long term, to economic inequalities and corrupting the role of higher education in society as a whole. For instance, in 2020 Ofqual, the UK Office of Qualifications and Examinations Regulation, used a MLDM to assess students for university admission decisions. Nearly 40% of students ended up receiving exam scores downgraded from their teachers' predictions, threatening to cost them their university spots. Analysis of the algorithm revealed that it had disproportionately hurt students from working-class and disadvantaged communities and inflated the scores of students from private schools [29].

[6]While the final acceptance decision is taken by humans, MLDMs are typically used as a first filter to "clean-up" the list from clear rejection cases.

[7]A platform that uses natural language processing and personality traits in order to help students find the suitable and right college for them.

[8]A Canadian startup that sells a cloud-based admissions assessment platform to over 300 schools.

Criminal risk assessment: There is an increasing adoption of MLDMs that predict risk scores based on historical data with the objective to guide human judges in their decisions. The most common use case is to predict whether a defendant will re-offend (or recidivate). Examples of risk assessment MLDMs include COMPAS [30], PSA [31], SAVRY [32], predPol [33]. Predicting risk and recidivism requires input information such as: number of arrests, type of crime, address, employment status, marital status, income, age, housing status, etc. Unfair risk assessment MLDMs, as revealed by the highly publicized 2016 proPublica article [34], may result in biased treatment of individuals based solely on their race. In extreme cases, it may lead to wrongful imprisonments for innocent people, contributing to the cycle of violation and crime.

Teachers evaluation and promotion: MLDMs are increasingly used by decision makers to decide which teachers to retain after a probationary period [35] and which tenured teachers to promote. An example of such MLDM is IMPACT [36]. Teacher evaluation MLDMs take as input teacher related features (age, education level, experience, surveys, classroom observations), students related features (test scores, sociodemographics, surveys), and principals related features (surveys about the school and teachers), to predict whether teachers are retained. A biased teacher evaluation MLDM may lead to a systematic unfair low evaluation for teachers in poor neighborhoods, which, very often, happen to be teachers belonging to minority groups [37]. On the long term, this may lead to a significant drop in students' performance and the compromise of overall school reputation [2].

Child maltreatment prediction: The objective of the MLDM in child maltreatment prediction is to estimate the likelihood of substantiated maltreatment (neglect, physical abuse, sexual abuse, or emotional maltreatment) among children. The system generates risk scores, which would then trigger a targeted early intervention in order to prevent children maltreatment. PRM (predictive risk model) [38] has been developed to estimate the likelihood of substantiated maltreatment among children enrolled in New Zealand's public benefit system. In Finland, the government uses a ML-based system called "Kela" to administer benefits and to identify risk factors indicating that a child might need welfare services. In the US, the Allegheny County uses AFST (Allegheny Family Screening Tool) [39] to improve decision-making in child welfare system. The features considered in this type of MLDM include both contemporaneous and historical information for children and caregivers. An unfair MLDM may use a proxy variable to predict decisions based on the community rather than which child gets harmed. For example, a major cause of unfairness in AFST is the rate of referral calls; the community calls the child abuse hotline to report non-white families at a much

higher rate than it does to report white families [39]. On the long term, this creates a vicious cycle as families which have been reported will be the subject of more scrutiny and more requirements to satisfy, and eventually, will be more likely to fail short of these requirements and hence confirm the prediction of the system.

Health care: Since decades, ML algorithms are able to process anonymized electronic health records and flag potential emergencies, to which clinicians are invited to respond promptly. Examples of features that might be used in disease (chronic conditions) prediction include vital signs, blood test, socio-demographics, education, health insurance, home ownership, age, race, address. The outcome of the MLDM is typically an estimated likelihood of getting a disease. A biased disease prediction MLDM can misclassify individuals in certain sub-populations in a disproportionately higher rate than the dominant population. For instance, diabetic patients have known differences in associated complications across ethnicities [40]. Obemeyer et al. [41] give another example of an MLDM that predicts the health care spending for individuals in the coming years (useful information for insurance companies). They observe that the MLDM is biased against African-Americans because it uses the cost of health services in the previous year to predict the spending in the coming years. As African-Americans were spending less on health services than whites in the previous year, they were predicted to be spending less in the coming years. Hence, for the same prediction score, African-Americans were found to be sicker (more health issues) than whites. Consequently, white patients were benefiting more from additional help programs than African-Americans. More generally, because different sub-populations might have different characteristics, a single model to predict complications is unlikely to be best-suited for specific groups in the population even if they are equally represented in the training data [42]. Failure to predict disease likelihood in a timely manner may, in extreme cases, have an impact on people's lives.

Online recommendation: Recommender systems are among the most widespread MLDMs in the market, with many services to assist users in finding products or information that are of potential interest [43]. Such systems find applications in various online platforms such as Amazon, Youtube, Netflix, LinkedIn, etc. An unfair recommender MLDM can amplify gender bias in the data. For example, a recommender MLDM called STEM, which aims to deliver advertisements promoting jobs in Science, Technology, Engineering, and Math fields, is deemed unfair as it has been shown that less women compared to men saw the advertisements due to gender imbalance [44]. Datta et al. [45] found that changing the gender bit in Google Ad Setting [46] resulted in a

significant difference in the type of job ads received: men received much more ads about high paying jobs and career coaching services towards high paying jobs compared to women.

Facial analysis: Automated facial analysis systems are used to identify perpetrators from security video footage, to detect melanoma (skin cancer) from face images [47], to detect emotions [48, 49, 50], and to even determine individual's characteristics such as IQ, propensity towards terrorist crime, etc. based on their face images [51]. The possible applications of Facial Analysis are innumerable. For instance, in France, FRT (Facial Recognition Tool) has been used on an experimental basis at various schools, with the aim of making access more fluid and secure for pupils. Furthermore, the government announced in 2020 that it would start to use an FRT system called "Alicem" in order to create a digital identification system by which its citizens could access governmental online services. Both of these, however, have sparked a lot of controversy leading to an announcement that the French government would be reviewing the use of FRT. Indeed, these devices are particularly intrusive and present major risks of invasion of the privacy and individual freedoms. Worse yet, a flawed MLDM may lead to biased outcomes such as wrongfully accusing individuals from specific ethnic groups (e.g. Asians, dark skin populations) for crimes (based on security video footage) at a much higher rate than the rest of the population. For instance, African-Americans have been reported to be more likely to be stopped and investigated by law enforcement due to a flawed face recognition system [52]. An investigation of three commercial face-based gender classification systems found that the error rate for dark-skinned females can be as high as 34.7% while for light-skinned males the maximum error rate is 0.8% [53].

Others: Other MLDMs with fairness concerns include: insurance policy prediction [54], income prediction [15], [55, 56, 57, 58], and university ranking [59, 2].

For a survey of the various kinds of MLDMs used in European countries, and a description of the debates and legal actions they have triggered, we recommend the excellent report by Robin Allen QC and Dee Masters [60] for the European Network of Equality Bodies.

4. Fairness notion selection criteria

In order to systemize the procedure for selecting the most suitable fairness notion for a specific MLDM system, we identify a set of criteria that can be used as as roadmap. For each criterion, we check whether it holds in the problem at hand or not. Telling whether a criterion is satisfied or not does not typically require an expertise in the problem domain.

This section presents a list of 13 selection criteria. These criteria are derived mainly from three sources. First, the types of bias. For instance, the unreliable outcome criterion is a manifestation of a historical bias. Second, the mathematical formulation of the fairness notions themselves. For instance, the emphasis on precision vs recall criterion reflects a fundamental difference in the mathematical formulations of two families of notions, namely, predictive parity and equal opportunity. Third, the existing anti-discrimination legislation. The last two criteria are inspired by the current legislation.

We note here that in some cases, these criteria can, not only indicate if a fairness notion is suitable, but whether it is "acceptable" to use in the first place.

Ground truth availability: A ground truth value is the true and correct *observed* outcome corresponding to given sample in the data. It should be distinguished from an *inferred* subjective outcome in historical data which is decided by a human. An example of a scenario where ground truth is available is when predicting whether an individual has a disease. The ground truth value is observed by submitting the individual to a blood test[9] for example. An example of a scenario where ground truth is not available is predicting whether a job applicant is hired. The outcome in the training data is inferred by a human decision maker which is often a subjective decision, no matter how hard she is trying to be objective. It is important to mention here that the availability of the ground truth depends on how the outcome is defined. Consider, for example, college admission scenario. If the outcome in the training data is defined as whether the applicant is admitted or rejected, ground truth is not available. If, however, the outcome is defined as whether the applicant will ultimately graduate from college with a high GPA, ground truth is available as it can be observed after a couple of years.

Base rate is the same across groups: The base rate is the proportion of positive outcome in a population (Table 1). A positive outcome is the goal of the prediction (e.g. a candidate to college is admitted, a child is maltreated, an individual is granted a loan, etc.). Note that the positive outcome can be desirable (e.g. hiring, admission) or undesirable (e.g. firing, high criminal risk). The base rate can be the same or differs across sub-populations. For example, the base rates for diabetes disease occurrence for men and women is typically the same. But, for another disease such as prostate cancer, the base rates are different between men and women[10].

[9] Assuming the blood test is flawless.

[10] While male prostate cancer is the second most common cancer in men, female prostate cancer is rare [61].

(Un)reliable outcome: In scenarios where ground truth is not available, the outcome (label) in the data is typically inferred by humans. The outcome in the training data in that case can or cannot be reliable as it can encode human bias. The reliability of the outcome depends on the data collection procedure and how rigorous the data has been checked. Scenarios such as job hiring and college admission may be more prone to the unreliable outcome problem than recommender system for example. A "one-size-fit-all" MLDM model in disease prediction that does not take into consideration the ethnic group of the individual may result in unreliable outcome as well.

Presence of explanatory variables: An explanatory variable[11] is correlated with the sensitive attribute (e.g. race) in a legitimate way. Any discrimination that can be explained using that variable is considered legitimate and is acceptable. For instance, if all the discrepancy between male and female job hiring rates is explained by their education levels, the discrimination can be deemed legitimate and acceptable.

Emphasis on precision vs recall: Precision (the complement of target population error [62]) is defined as the fraction of positive instances among the predicted positive instances. In other words, if the system predicts an instance as positive, how precise that prediction is. Recall (the complement of model error [62]) is defined as the fraction of the total number of positive instances that are correctly predicted positive. In other words, how many of the positive instances the system is able to identify. There is always a trade-off between precision and recall (increasing one will lead, very often, to decreasing the other). Depending on the scenario at hand, the fairness of the MLDM may be more sensitive to one on the expense of the other. For example, granting loans to the maximum number of deserving applicants contributes more to fairness than making sure that an applicant who has been granted a loan really deserves it[12]. When firing employees, however, the opposite is true: fairness is more sensitive to wrongly firing an employee, rather than, firing the maximum number of under-performing employees.

Emphasis on false positive vs false negative: Fairness can be more sensitive to false positive misclassification (type I error) rather than false negative misclassification (type II error), or the opposite. For example, in criminal risk assessment scenario, it is commonly accepted that

[11] Referred also as a resolving variable.

[12] It is important to mention here that from the loan granting organization's point of view, the opposite is true. That is, it is more important to make sure that an applicant who has been granted a loan really deserves it and will not default in payments because the interest payments resulting from a loan are relatively small compared to the loan amount that could be lost. Our aim here is fairness, while the loan granting organization's goal is benefit.

incarcerating an innocent person (false positive) is more serious than letting a guilty person escape (false negative).

Cost of misclassification: Depending on the scenario at hand, the cost of misclassification can be significant (e.g. incarcerating an individual, firing an employee, rejecting a college application, etc.) or mild and without consequential impact (e.g. useless product recommendation, misleading income prediction, offensive online translation, abusive results in online autocomplete, etc.)

Prediction threshold is fixed or floating: Decisions in MLDM are typically made based on predicted real-valued score. In the case of binary outcome, the score is turned into a binary value such as $\{0, 1\}$ by thresholding[13]. In some scenarios, it is desirable to interpret the real-value score as probability of being accepted (predicted positive). The threshold used as a cutoff point where positive decisions are demarcated from negative decisions can be fixed or floating. A fixed threshold is set carefully and tends to be valid for different datasets and use cases. For instance, in recidivism risk assessment, high risk threshold is typically fixed. A floating threshold can be selected and fine-tuned arbitrarily by practitioners to accommodate a changing context. Acceptance score in loan granting scenarios is an example of a floating threshold as it can move up or down depending on the economic context. When the threshold is floating in a given application, assessing fairness should be done using a suitable fairness notion (e.g. calibration) otherwise, the result of the assessment may be misleading for specific threshold values.

Likelihood of intersectionality: Intersectionality theory [63] focuses on a specific type of bias due to the combination of sensitive factors. An individual might not be discriminated based on race only or based on gender only, but she might be discriminated because of a combination of both. Black women are particularly prone to this type of discrimination.

Likelihood of masking: Masking is a form of intentional discrimination that allows decision makers with prejudicial views to mask their intentions [64]. Masking is typically achieved by exploiting how fairness notions are defined. For example, if the fairness notion requires equal number of candidates to be accepted from two ethnic groups, the MLDM can be designed to carefully select candidates from the first group (satisfying strict requirements) while selecting randomly from the second group just to "make the numbers".

Sources of Bias: Bias in the MLDM outcome can arise from several possible sources at any

[13]The threshold is defined by the decision makers depending on the context of interest.

stage in the data generation and machine learning pipeline. Framing sources of bias necessitates deep understanding of the application at hand and, typically, can only be identified after a "post-mortem" analysis of the predicted outcome. However, in some real-world scenarios, one or more sources of bias may be more likely than others. In such cases, the suspected source of bias can be used as a criterion to select the most appropriate notion for fairness assessment. Sources of bias can be grouped broadly into six categories: historical, representation, measurement, aggregation, evaluation, and deployment [42]. Historical bias arises when the data reliably collected from the world leads to outcomes which are unwanted and socially unfavorable. For example, while data reliable collected indicates that only 5% of Fortune 500 CEOs are women [65], the resulting outcome of a prediction system based on this data is typically not wanted[14]. Representation bias arises when some non-protected populations are under-represented in the training data. Measurement bias arises when the features or label values are not measured accurately. For example, Street Bump is an application used in Boston city to detect when residents drive over potholes thanks to the accelerometers built into smartphones [66]. Collecting data using this application introduces a measurement bias due to the disparity in the distribution of smartphones according to the different districts in the city, which are often correlated with race or level of income. Aggregation bias arises when sub-populations are aggregated together while a single model is unlikely to fit all sub-populations. For instance, the genetic risk scores derived largely on European populations have been shown to generally perform very poorly in the prediction of osteoporotic fracture and bone mineral density on non-European populations, in particular, on Chinese population [67]. Evaluation bias arises when the training data differs significantly from the testing data. For instance, several MLDMs are trained using benchmark datasets which may be very different from the target dataset. Deployment bias arises when there is a disparity between the initial purpose of an MLDM and the way it is actually used. For instance, a child maltreatment MLDM might be designed to predict the risk of child abuse after two years from the reception of a referral call, while in practice it may be used to help social agents take decisions about an intervention. This can lead to a bias since the decision has an impact on the outcome [23].

Legal Framework: Anti-discrimination regulations in several countries, in particular US, distinguish between two legal frameworks, namely disparate treatment and disparate impact [64]. In

[14]For this reason, Google has changed their image search result for CEO to return a higher proportion of women.

the disparate treatment framework, a decision is considered unfair if it uses (directly or indirectly) the individual's sensitive attribute information. In the disparate impact framework, a decision is unfair if it results in an outcome that is disproportionately disadvantageous (or beneficial) to individuals according to their sensitive attribute information. Zafar et al. [5] formalized another fairness criterion, namely, disparate mistreatment according to which, a decision is unfair if it results in different misclassification rates for groups of people with different sensitive attribute information. Note that this criterion is currently not supported by a legal framework. Machine learning fairness notions can be classified according to the type of fairness it is evaluating. For instance, if a plaintiff is accusing an employer for intentional discrimination, she should consider the disparate treatment legal framework, and hence a fairness notion which falls in that framework.

The existence of regulations and standards: In some domains, laws and regulations might be imposed to avoid discrimination and bias. For instance, guidelines from the *U.S. Equal Employment Opportunity Commission* state that a difference of the probability of acceptance between two sub-populations exceeding 20% is illegal [8]. Another example might be an internal organizational policy imposing diversity among its employees.

5. Fairness notions

Let V, A, and X be three random variables representing, respectively, the total set of attributes, the sensitive attributes, and the remaining attributes describing an individual such that $V = (X, A)$ and $P(V = v_i)$ represents the probability of drawing an individual with a vector of values v_i from the population. For simplicity, we focus on the case where A is a binary random variable where $A = 0$ designates the protected group, while $A = 1$ designates the non-protected group. Let Y represent the actual outcome and $\hat{Y}$ represent the outcome returned by the prediction algorithm (MLDM). 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. A perfect MLDM will match perfectly the actual outcome ($\hat{Y} = Y$). Typically, the predicted outcome $\hat{Y}$ is derived from a score represented by a random variable S where $P(S = s)$ is the probability that the score value is equal to s.

All fairness notions presented in this section address the following question: "is the outcome/prediction of the MLDM fair towards individuals?". So fairness notion is defined as a mathematical condition that must involve either $\hat{Y}$ or S along with the other random variables. As

such, we are not concerned by the inner-workings of the MLDM and their fairness implications. What matters is only the score/prediction value and how fair/biased it is.

Most of the proposed fairness notions are properties of the joint distribution of the above random variables (X, A, Y, $\hat{Y}$, and S). They can also be interpreted using the confusion matrix and the related metrics (Table 1).

Table 1: Metrics based on confusion matrix.

	Actual Positive $Y = 1$	Actual Negative $Y = 0$		
Predicted Positive $\hat{Y} = 1$	**TP** (True Positive)	**FP** (False Positive) *Type I error*	**PPV** = $\frac{TP}{TP+FP}$ *Positive Predictive Value* *Precision* *PV+* *Target Population Error*	**FDR** = $\frac{FP}{TP+FP}$ *False Discovery Rate* *Target Population Error*
Predicted Negative $\hat{Y} = 0$	**FN** (False Negative) *Type II error*	**TN** (True Negative)	**FOR** = $\frac{FN}{FN+TN}$ *False Omission Rate* *Success Predictive Error*	**NPV** = $\frac{TN}{FN+TN}$ *Negative Predictive Value* *PV-*
	TPR = $\frac{TP}{TP+FN}$ *True Positive Rate* *Sensitivity* *Recall*	**FPR** = $\frac{FP}{FP+TN}$ *False Positive Rate* *Model Error*	**OA** = $\frac{TP+TN}{TP+FP+TN+FN}$ *Overall Accuracy*	**BR** = $\frac{TP+FN}{TP+FP+TN+FN}$ *Base Rate* *Prevalence (p)*
	FNR = $\frac{FN}{TP+FN}$ *False Negative Rate* *Model Error*	**TNR** = $\frac{TN}{FP+TN}$ *True Negative Rate* *Specificity*		

While presenting and discussing fairness notions, whenever needed, we use the simple job hiring scenario of Table 2. 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 gender, that is, we are focusing on whether male and female applicants are treated equally. Table 2(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.

Table 2: 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						(b) Prediction	
Gender	Education Level	Job Experience	Age	Marital Status	Y	$\hat{Y}$	S
Female 1	8	2	39	single	0	1	0.5
Female 2	8	2	26	married	1	0	0.1
Female 3	12	8	32	married	1	1	0.5
Female 4	11	3	35	single	0	0	0.2
Female 5	9	5	29	married	1	0	0.3
Male 1	11	3	34	single	1	1	0.8
Male 2	8	0	48	married	0	0	0.1
Male 3	7	3	43	single	1	0	0.1
Male 4	8	2	26	married	1	1	0.5
Male 5	8	2	41	single	0	1	0.5
Male 6	12	8	30	single	1	1	0.8
Male 7	10	2	28	married	1	0	0.3

A simple and straightforward approach to address fairness problem is to ignore completely any sensitive attribute while training the MLDM system. This is called *fairness through unawareness*[15]. We don't treat this approach as fairness notion since, given MLDM prediction, it does not allow to tell if the MLDM is fair or not. Besides, it suffers from the basic problem of proxies. Many attributes (e.g. home address, neighborhood, attended college) might be highly correlated to the sensitive attributes (e.g. race) and act as proxies of these attributes. Consequently, in almost all situations, removing the sensitive attribute during the training process does not address the problem of fairness.

5.1. Statistical parity

Statistical parity [70] (a.k.a demographic parity [71], independence [72], equal acceptance rate [73], benchmarking [74], group fairness [70]) is one of the most commonly accepted notions of fairness. It requires the prediction to be statistically independent of the sensitive attribute ($\hat{Y} \perp A$). Thus, a

[15]Known also as: blindness, unawareness [3], anti-classification [68], and treatment parity [69].

classifier $\hat{Y}$ satisfies statistical parity if:

$$P(\hat{Y} \mid A = 0) = P(\hat{Y} \mid A = 1) \tag{1}$$

In other words, the predicted acceptance rates for both protected and unprotected groups should be equal. Using the confusion matrix (Table 1), statistical parity implies that $(TP + FP)/(TP + FP + FN + TN)$ should be equal for both groups. In the MLDM of Table 2, it means that one should not hire proportionally more applicants from one group than the other. 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, the MLDM of Table 2 does not satisfy statistical parity.

Statistical parity is appealing in scenarios where there is a preferred decision over the other, and provided there are no other considerations relevant for the decision, in which case, the following fairness notion namely, conditional statistical parity, is more suitable. For example, being accepted to a job, not being arrested, being admitted to a college, etc.[16]. What really matters is a balance in the prediction rate among all groups.

Statistical parity is suitable when the label Y is not trustworthy due to some flawed or biased measurement[17]. An example of this type of problem was observed in the recidivism risk prediction tool COMPAS [34]. Because minority groups are more controlled, and more officers are dispatched in their regions, the number of arrests (used to assess the level of crime [42]) of those minority groups is significantly higher than that of the rest of the population. Hence, for fairness purposes, in the absence of information to precisely quantify the differences in recidivism by race, the most suitable approach is to treat all sub-populations equally with respect to recidivism [76].

Statistical parity is also well adapted to contexts in which some regulations or standards are imposed. For example, a law might impose to equally hire or admit applicants from different sub-populations.

The main problem of statistical parity is that it doesn't consider a potential correlation between the label Y and the sensitive attribute A. In other words, if the underlying base rates of the protected and unprotected groups are different, statistical parity will be misleading. In particular, modifying

[16]This might not be the case in other scenarios such as disease prediction, child maltreatment, where imposing a parity of positive predictions is meaningless.

[17]This is also known as differential measurement error [75].

an MLDM with a perfect prediction ($\hat{y} = y$) so to satisfy statistical parity while the base rates are different will lead to loss of utility [77]. As an example, Figure 1 illustrates a scenario for hiring computer engineers where equal proportions of male/female applicants have been predicted hired (60%) thus, satisfying statistical parity. However, when considering the label and more precisely the base rates that differ in both groups (0.3 for men versus 0.4 for women), the classifier becomes discriminative against female applicants (50% of qualified female applicants are not predicted hired). More generally, when the ground truth is available and is used in the training of the MLDM, statistical parity is not recommended because, very often, it conflicts with the ground truth [5].

Another issue with this notion is its "laziness"; if we hire carefully selected applicants from male group and random applicants from female group, we can still achieve statistical parity, yet leading to negative results for the female group as its performance will tend to be worse than that of male group. This practice is an example of *self-fulfilling prophecy* [70] where a decision maker may simply select random members of a protected group rather than qualified ones, and hence, intentionally building a bad track record for that group. Barocas and Selbst refer to this problem as masking [64]. Masking is possible to game several fairness notions, but it is particularly easy to carry out in the case of statistical parity.

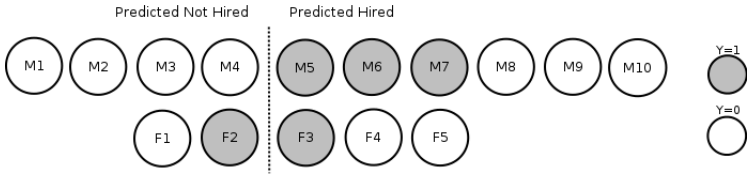

Figure 1: F_i and M_i ($i \in [1-10]$) designate female and male applicants, respectively. The grey shaded circles indicate applicants who belong to the positive class while white circles indicate applicants belonging to the negative class. The dotted vertical line is the prediction boundary. Thus, applicants at the right of this line are predicted hired while applicants at the left are predicted not hired.

5.2. Conditional statistical parity

Conditional statistical parity [7], called also conditional discrimination-aware classification in [78] is a variant of statistical parity obtained by controlling on a set of legitimate attributes[18]. The legitimate attributes (we refer to them as E) among X are correlated with the sensitive attribute A and give some factual information about the label at the same time leading to a *legitimate*

[18]Called explanatory attributes in [78].

discrimination. In other words, this notion removes the illegal discrimination, allowing the disparity in decisions to be present as long as they are explainable [7]. In the hiring example, possible explanatory factors that might affect the hiring decision for an applicant could be the education level and/or the job experience. If the data is composed of many highly educated and experienced male applicants and only few highly educated and experienced women, one might justify the disparity between predicted acceptance rates between both groups and consequently, does not necessarily reflect gender discrimination. Conditional statistical parity holds if:

$$P(\hat{Y} = 1 \mid E = e, A = 0) = P(\hat{Y} = 1 \mid E = e, A = 1) \quad \forall e \tag{2}$$

Table 3: Application of conditional statistical parity by controlling on education level and job experience.

(a) Dataset						(b) Prediction	
Gender	Education Level	Job Experience	Age	Marital Status	Y	$\hat{Y}$	S
Female 1	8	2	39	single	0	1	0.5
Female 2	8	2	26	married	1	0	0.1
Female 3	12	8	32	married	1	1	0.5
Male 4	8	2	26	married	1	1	0.5
Male 5	8	2	41	single	0	1	0.5
Male 6	12	8	30	single	1	1	0.8

Table 3 shows two possible combinations values for E. The first combination (education level=8 and job experience=2) includes samples Female 1, Female 2, Male 4, and Male 5 for which the prediction is clearly discriminative against women as the predicted acceptance rates for men and women are 1 and 0.5, respectively. The second combination (education level=12 and job experience=8) includes Female 3 and Male 6 in which the prediction is fair (predicted acceptance rate is 1 for both applicants). Overall, the prediction is not fair as it does not hold for one combination of values of E.

In practice, conditional statistical parity is suitable when there is one or several attributes that justify a possible disparate treatment between different groups in the population. Hence, choosing the legitimate attribute(s) is a very sensitive issue as it has a direct impact on the fairness of the decision-making process. More seriously, conditional statistical parity gives a decision maker a tool to game the system and realize a self-fullfilling prophecy. Therefore, it is recommended to resort to

domain experts or law officers to decide what is unfair and what is tolerable to use as legitimate discrimination attribute [78].

5.3. Equalized odds

Unlike the two previous notions, equalized odds [79] (separation in [72], conditional procedure accuracy equality in [12], disparate mistreatment in [5], error rate balance in [9]) considers both the predicted and the actual outcomes. Thus, the prediction is conditionally independent from the protected attribute, given the actual outcome ($\hat{Y} \perp A \mid Y$). In other words, equalized odds requires both sub-populations to have the same TPR and FPR (Table 1). In our example, this means that the probability of an applicant who is actually hired to be predicted hired and the probability of an applicant who is actually not hired to be incorrectly predicted hired should be both same for men and women:

$$P(\hat{Y} = 1 \mid Y = y,\ A = 0) = P(\hat{Y} = 1 \mid Y = y,\ A = 1) \quad \forall y \in \{0, 1\} \tag{3}$$

In the example of Table 2, the TPR for male and female groups is 0.6 and 0.33, respectively while the FPR is exactly the same (0.5) for both groups. Consequently, the equalized odds does not hold.

By contrast to statistical parity, equalized odds is well-suited for scenarios where the ground truth exists such as: disease prediction or stop-and-frisk [80]. It is also suitable when the emphasis is on recall (the fraction of the total number of positive instances that are correctly predicted positive) rather than precision (making sure that a predicted positive instance is actually a positive instance).

A potential problem of equalized odds is that it may not help closing the gap between the protected and unprotected groups. For example, consider a group of 20 male applicants of which 16 are qualified and another equal size group of 20 females of which only 2 are qualified. If the employer decides to hire 9 applicants and while satisfying equalized odds, 8 offers will be granted to the male group and only 1 offer will be granted to the female group. While this decision scheme looks fair on the short term, on the long term, however, it will contribute to confirm this "unfair" status-quo and perpetuate this vicious cycle[19]. Whether to consider this long term impact as a problem of

[19] If the job is a well-paid, male group tends to have a better living condition and affords better education for their kids, and thus enable them to be qualified for such well-paid jobs when they grow up. The gap between the groups will tend to increase over time.

equalized odds is a controversial issue as it overlaps with the different but related question of "how to address unfairness?". Note that other fairness notions, such as statistical parity, help closing the gap between the protected and unprotected groups on the long term.

Because equalized odds requirement is rarely satisfied in practice, two variants can be obtained by relaxing Eq. 3. The first one is called **equal opportunity** [79] (false negative error rate balance in [9]) and is obtained by requiring only TPR equality among groups:

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1) \tag{4}$$

In the job hiring example, this is to say that we should hire equal proportion of individuals from the qualified fraction of each group.

As $TPR = TP/(TP + FN)$ (Table 1) does not take into consideration FP, equal opportunity is completely insensitive to the number of false positives. This is an important criterion when considering this fairness notion in practice. More precisely, in scenarios where a disproportionate number of false positives among groups has fairness implications, equal opportunity should not be considered. The scenario in Table 4 shows an extreme case of a job hiring dataset where the male group has a large number of false positives (Male $7 - 100$) while equal opportunity is satisfied.

Table 4: An extreme job hiring scenario satisfying equal opportunity. All Male $7 - 100$ samples are false positives (label Y is 0 and prediction $\hat{Y}$ is 1).

(a) Dataset						(b) Prediction	
Gender	Education Level	Job Experience	Age	Marital Status	Y	$\hat{Y}$	S
Female 1	8	2	39	single	1	1	0.5
Female 2	8	2	26	married	0	0	0.1
Female 3	12	8	32	married	1	0	0.3
Male 4	8	2	26	married	1	1	0.5
Male 5	8	2	41	single	0	0	0.2
Male 6	12	8	30	single	1	0	0.4
Male 7	10	5	32	married	0	1	0.8
...	...	...	...	...	0	1	...
Male 100	8	10	27	single	0	1	0.7

To decide about the suitability of equal opportunity in the job hiring example, the question that should be answered by stakeholders and decision makers is "if all other things are equal, is it fair to

hire disproportionately more unqualified male candidates?". For the employer, it is undesirable to have several false positives (regardless of their gender) as the company will end up with unqualified employees. For a stakeholder whose goal is to guarantee fairness between males and females, it is not very critical to have more false positives in one group, provided that these two groups have the same proportion of false negatives (a qualified candidate which is not hired).

In the scenario of predicting which employees to fire, however, a false positive (firing a well-performing employee) is critical for fairness. Hence, equal opportunity should not be used as a measure of fairness.

The second relaxed variant of equalized odds is called **predictive equality** [7] (false positive error rate balance in [9]) which requires only the FPR to be equal in both groups.

In other words, predictive equality checks whether the accuracy of decisions is equal across protected and unprotected groups:

$$P(\hat{Y} = 1 \mid Y = 0, A = 0) = P(\hat{Y} = 1 \mid Y = 0, A = 1) \tag{5}$$

In the job hiring example, predictive equality holds when the probability of an applicant with an actual weak profile for the job to be incorrectly predicted hired is the same for both men and women.

Since $FPR = FP/(FP+TN)$ (Table 1) is independent from FN, predictive equality is completely insensitive to false negatives. One can come up with an extreme example similar to Table 4 with a disproportionate number of false negatives but predictive equality will still be satisfied (keeping all other rates equal). Hence, in scenarios where fairness between groups is sensitive to false negatives, predictive equality should not be used. Such scenarios include hiring and admission where a false negative means a qualified candidates are rejected disproportionately among groups. Predictive equality is acceptable in criminal risk assessment scenarios as false negatives (releasing a guilty person) are less critical than false positives (incarcerating an innocent person).

Predictive equality is particularly suitable to measure the fairness of face recognition systems in crime investigation where security camera footage are analyzed. Fairness between ethnic groups with distinctive face features is very sensitive to the FPR. A false positive means an innocent person is being flagged as participating in a crime. If this false identification happens at a much higher rate for a specific sub-population (e.g. dark skinned group) compared to the rest of the population, it is clearly unfair for individuals belonging to that sub-population.

Looking to the problem from another perspective, choosing between equal opportunity and predictive equality depends on how the outcome/label is defined. In scenarios where the positive outcome is desirable (e.g. hiring, admission), typically fairness is more sensitive to false negatives rather than false positives, and hence equal opportunity is more suitable. In scenarios where the positive outcome is undesirable for the subjects (e.g. firing, risk assessment), typically fairness is more sensitive to false positives rather than false negatives, and hence predictive equality is more suitable.

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

Proposition 5.1. *Satisfying equal opportunity and predictive equality is equivalent to satisfying equalized odds:*

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

5.4. Conditional use accuracy equality

Conditional use accuracy equality [12] (called sufficiency in [72]) is achieved when all population groups have equal $PPV = \frac{TP}{TP+FP}$ and $NPV = \frac{TN}{FN+TN}$. 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:

$$P(Y = y \mid \hat{Y} = y, A = 0) = P(Y = y \mid \hat{Y} = y, A = 1) \quad \forall y \in \{0, 1\} \tag{6}$$

Intuitively, this definition implies equivalent accuracy for male and female applicants from both positive and negative predicted classes [13]. By contrast to equalized odds (Section 5.3), one is conditioning on the algorithm's predicted outcome not the actual outcome. In other words, this notion emphasis the precision of the MLDM system rather than its sensitivity (a trade-off discussed earlier in Section 4).

The calculated PPVs for male and female applicants in our hiring example (Table 2) are 0.75 and 0.5, respectively. NPVs for male and female applicants are both equal to 0.33. Overall the dataset in Table 2 does not satisfy conditional use accuracy equality.

Predictive parity [9] (called outcome test in [74]) is a relaxation of conditional use accuracy equality requiring only equal PPV among groups:

$$P(Y = 1 \mid \hat{Y} = 1, A = 0) = P(Y = 1 \mid \hat{Y} = 1, A = 1) \tag{7}$$

In our example, this is to say that the prediction used to determine the candidate's eligibility for a particular job should reflect the candidate's actual capability of doing this job which is harmonious with the employer's benefit.

Like predictive equality (Eq. 5), predictive parity is insensitive to false negatives. Hence in any scenario where fairness is sensitive to false negatives, predictive parity should not be considered sufficient.

Choosing between predictive parity and equal opportunity depends on whether the scenario at hand is more sensitive to precision or recall. For precision-sensitive scenarios, typically predictive parity is more suitable while for recall-sensitive scenarios, equal opportunity is more suitable. Precision-sensitive scenarios include disease prediction, child maltreatment risk assessment, and firing from jobs. Recall-sensitive scenarios include loan granting, recommendation systems, and hiring. Very often, precision-sensitive scenarios coincide with situations where the positive prediction ($\hat{Y} = 1$) entails a higher cost [5]. For example, a predicted child maltreatment case will result in placing the child in a foster house which will generally entail a higher cost compared to a negative prediction (low risk of child maltreatment) in which case the child stays with the family and typically no action is taken.

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 notion and give it the name **negative predictive parity**.

Definition 5.1. *Negative predictive parity holds iff all sub-groups have the same* $NPV = \frac{TN}{FN+TN}$*:*

$$P(Y = 1 \mid \hat{Y} = 0, A = 0) = P(Y = 1 \mid \hat{Y} = 0, A = 1) \tag{8}$$

The following proposition states formally the relationship between conditional use accuracy equality, predictive parity, and negative predictive parity.

Proposition 5.2. *Satisfying predictive parity and negative predictive parity is equivalent to satisfying*

conditional use accuracy equality:

$$Eq.\ 6 \Leftrightarrow Eq.\ 7 \wedge Eq.\ 8$$

5.5. Overall accuracy equality

Overall accuracy equality [12] is achieved when overall accuracy for both groups is the same. Thus, true negatives and true positives are equally considered and desired. Using the confusion matrix (Table 1), this implies that $(TP + TN)/(TP + FN + FP + TN)$ is equal for both groups. In our example, it is to say that the probability of well-qualified applicants to be correctly accepted for the job and non-qualified applicants to be correctly rejected is the same for both male and female applicants:

$$P(\hat{Y} = Y|A = 0) = P(\hat{Y} = Y|A = 1) \tag{9}$$

Table 5: A job hiring scenario satisfying overall accuracy but not conditional use accuracy equality.

	Group 1 (Female)			Group 2 (Male)			
	Gender	Y	$\hat{Y}$	Gender	Y	$\hat{Y}$	
	F1	1	1	M1	1	1	
	F2	1	0	M2	0	1	
OA = 0.625	F3	1	0	M3	0	1	OA = 0.625
PPV = 1	F4	0	0	M4	0	0	PPV = 0.4
NPV = 0.25	F5	1	1	M5	0	0	NPV = 1
	F6	1	1	M6	0	0	
	F7	1	0	M7	0	1	
	F8	1	1	M8	1	1	

Overall accuracy equality is closely related to equalized odds (Eq. 3) and to conditional use accuracy equality (Eq. 6). The main difference is that overall accuracy equality aggregates together positive class and negative class misclassifications (FP and FN). Aggregating together FP and FN (and hence TP and TN) without any distinction is very often misleading for fairness purposes.

Proposition 5.3. *An MLDM that satisfies equalized odds or conditional use accuracy equality always satisfies overall accuracy.*

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

The reverse, however, is not true. That is, an MLDM that satisfies overall accuracy does not necessarily satisfy equalized odds or conditional use accuracy equality. To prove it, consider the example in Table 5 satisfying overall accuracy equality but not conditional use accuracy equality. For the female group, there are only FN misclassifications (no FP) and more TPs than TNs, while in the male group, there are only FP misclassifications (no FN) and more TNs than TPs. But since the proportion of correct classifications is the same in both groups (5 out of 8), overall accuracy equality holds. In real-world applications, it is very uncommon that TP (or FN) and TN (or FP) are desired at the same time and without distinction. For example, overall accuracy equality is not suitable to measure fairness in child maltreatment prediction because a False Positive (misclassifying a child case which is not at risk[20]) is less damaging than a False Negative (misclassifying a child case which is at risk[21]). A hypothetical health care scenario where overall accuracy equality is suitable is when both types of misclassifications have the same cost/benefit. For example, an eventual health condition that yields very similar complications (1) when the treatment is administered wrongly and (2) when the treatment is not administered while it is needed.

5.6. Treatment equality

Treatment equality [12] is achieved when the ratio of FPs and FNs is the same for both protected and unprotected groups:

$$\frac{FN}{FP}^{(\mathrm{A}=0)} = \frac{FN}{FP}^{(\mathrm{A}=1)} \tag{10}$$

Treatment equality is insensitive to the numbers of TPs and TNs which are important to identify bias between sub-populations in most real-world scenarios. Berk et al. [12] note that treatment equality can serve as an indicator to achieve other kinds of fairness. Table 6 shows a dataset which fails to satisfy all previous notions, yet, treatment equality is satisfied. Treatment equality can be used in real-world scenarios where only the type of rate of misclassification matters for fairness.

Treatment equality can be suitable to use in case the cost (or benefit) of a FP is a fixed ratio (or reciprocal) of the cost (or benefit) of a FN. For example, one can think of a loan granting scenario where the cost of a FP (misclassifying a non-defaulter) is exactly a fraction (e.g. 1/3) of the cost of

[20] Results in a useless intervention, because the child is not at risk anyway.
[21] Results in a failure to anticipate a child maltreatment.

Table 6: A job hiring scenario satisfying treatment equality but not satisfying all of the previous notions.

			Group 1 (Female)			Group 2 (Male)					
			Gender	Y	$\hat{Y}$	Gender	Y	$\hat{Y}$			
			F1	1	1	M1	1	1			
TPR	=	0.33	F2	0	0	M2	1	1	TPR	=	0.8
FPR	=	0.8	F3	0	1	M3	1	1	FPR	=	0.66
PPV	=	0.2	F4	0	1	M4	1	1	PPV	=	0.66
NPV	=	0.33	F5	0	1	M5	0	0	NPV	=	0.5
OA	=	0.25	F6	0	1	M6	0	1	OA	=	0.625
FN/FP	=	0.5	F7	1	0	M7	0	1	FN/FP	=	0.5
			F8	1	0	M8	1	0			

a FN (misclassifying a defaulter).

Total fairness [12] is another notion which holds when all aforementioned fairness notions are satisfied simultaneously, that is, statistical parity, equalized odds, conditional use accuracy equality (hence, overall accuracy equality), and treatment equality. Total fairness is a very strong notion which is very difficult to hold in practice. Table 7 shows a scenario where total fairness holds. More generally, total fairness is satisfied in the very uncommon situation where the proportions of TPs, TNs, FPs, and FNs are the same in all groups.

Table 7: A job hiring scenario satisfying total fairness.

			Group 1 (Female)			Group 2 (Male)					
			Gender	Y	$\hat{Y}$	Gender	Y	$\hat{Y}$			
			F1	1	1	M1	1	1			
TPR	=	0.5	F2	0	0	M2	1	1	TPR	=	0.5
FPR	=	0.66	F3	0	1	M3	0	0	FPR	=	0.66
PPV	=	0.33	F4	0	1	M4	0	0	PPV	=	0.33
NPV	=	0.5	F5	1	0	M5	0	1	NPV	=	0.5
OA	=	0.4				M6	0	1	OA	=	0.4
FN/FP	=	0.5				M7	0	1	FN/FP	=	0.5
						M8	0	1			
						M9	1	0			
						M10	1	0			

Total fairness can be considered in scenarios where any deviation in misclassification or acceptance

rates between sub-populations is very costly[22].

5.7. Balance

The predicted outcome ($\hat{Y}$) is typically derived from a score (S) which is returned by the ML algorithm. All aforementioned fairness notions do not use the score to assess fairness. Typically, the score value is normalized to be in the interval $[0, 1]$ which makes it possible to interpret the score as the probability to predict the sample as positive. **Balance for positive class** [6] 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. In other words, a violation of this balance means that applicants belonging to the positive class in one group might receive steadily lower predicted score than applicants belonging to the positive class in the other group:

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

Table 8 shows a job hiring scenario where the average score for female candidates that should be hired ($Y = 1$) is 7.1 while it is 4.7 for male candidates. The scenario is not balanced for positive class. Note that, despite the significant difference between these two average values, for a score threshold value of 5, the scenario of Table 8 satisfies both statistical parity (Eq. 1) and equal opportunity (Eq. 4).

Table 8: A job hiring scenario satisfying statistical parity and equal opportunity (for a score threshold value of 5) but neither balance for positive class nor balance for negative class.

(a) Group 1 (Female)

Gender	Y	S
F1	1	9
F2	1	8
F3	0	8
F4	1	4.5
F5	0	4.5
F6	0	3.5

(b) Group 2 (Male)

Gender	Y	S
M1	1	6.2
M2	1	6
M3	0	5.5
M4	0	1
M5	1	2
M6	0	2

Balance of negative class [6] is an analogous fairness notion where the focus is on the negative

[22]The cost can be financial, ethical, reputation, etc.

class:

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

The scenario in Table 8 is not balanced for the negative class either since the average scores for the negative class ($Y = 0$) for the female and male groups are 5.3 and 2.8, respectively.

Both variants of balance can be required simultaneously (Eq. 11 and 12) which leads to a stronger notion of balance. Since no previous work reported such fairness notion, for completeness, we define it and call it **overall balance**.

Definition 5.2. *Overall balance is satisfied iff:*

$$E[S \mid Y = y, A = 0] = E[S \mid Y = y, A = 1] \quad \forall y \in \{0, 1\} \tag{13}$$

Balance fairness notions are relevant in the criminal risk assessment scenario because a divergence in the score values of individuals from different races may indicate a difference in the type of crime that can be committed (high risk score typically means a serious crime). Balance fairness notions are also suitable in the teacher firing scenario since any discrepancy between the average evaluation scores of fired teachers in different groups is a clear indicator of bias. On the other hand, balance fairness notions can be misleading in presence of clusters of samples sharing very similar attribute values and having score values in the vicinity of the positive/negative outcome threshold. In such case, the average score of the positive/negative class can change significantly due to a slight increase/decrease of the threshold value.

5.8. Calibration

Calibration [9] (a.k.a. test-fairness [9], matching conditional frequencies [79]) relies on the score variable as follows. To satisfy calibration, for each predicted probability score $S = s$, individuals in all groups should have the same probability to actually belong to the positive class:

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

Eq. 14 is very unlikely to be satisfied in practice as the probability of two individuals having exactly the same real number score is very small. Moreover, technically, the probability that S

exactly equal to s is typically 0. Therefore, in practice, the space of score values $[0, 1]$ is binned into intervals called bins such that any two values falling in the same bin are considered equal [6, 13, 81].

In our job hiring example, this implies that for any score value $s \in [0, 1]$, the probability of truly being hired should be the same for both male and female applicants.

Table 9: A job hiring scenario satisfying predictive parity (for any threshold smaller than 0.7 or larger than 0.8) but not calibration.

(a) Group 1 (Female)

Gender	Y	S
F1	1	0.85
F2	1	0.8
F3	0	0.8
F4	1	0.7
F5	0	0.7
F6	0	0.4
F7	1	0.4
F8	0	0.4

(b) Group 2 (Male)

Gender	Y	S
M1	1	0.85
M2	1	0.8
M3	1	0.8
M4	0	0.7
M5	0	0.7
M6	1	0.4
M7	0	0.4
M8	0	0.4

Eq. 14 is very similar to Eq. 7 corresponding to predictive parity. Table 9 illustrates a job hiring scenario that may or may not satisfy predictive parity depending on the score threshold to hire a candidate; for a threshold value of 0.6, PPV rate for both male and female groups is the same, 0.6, while for a threshold value of 0.75, PPV for female group is 0.66 but for male it is 1.0. However, the calibration score ($P(Y = 1 \mid S = s, A = a)$ $a \in \{0, 1\}$, $s \in [0, 1]$) for every value of s is as follows:

s	0.4	0.7	0.8	0.85
Female	0.33	0.5	0.5	1.0
Male	0.33	0	1.0	1.0

Calibration is satisfied for score values 0.4 and 0.85, but not satisfied for score values 0.7 and 0.8. Overall, the scenario of Table 9 does not satisfy calibration.

Interestingly, calibration is not always stronger than predictive parity [82]. Table 10 shows a job hiring scenario satisfying calibration, but not predictive parity. Calibration is suitable to use in scenarios where the threshold is not fixed and is very likely to be tuned to accommodate a changing context. A first example is the acceptance score in loan granting applications which may change abruptly due to economic instability. A second example is the child maltreatment risk assessment

Table 10: A job hiring scenario satisfying calibration but not predictive parity (for any threshold).

(a) Group 1 (Female)

Gender	Y	S
F1	1	0.8
F2	1	0.8
F3	1	0.7
F4	1	0.7
F5	0	0.7
F6	0	0.7
F7	0	0.3
F8	0	0.3

(b) Group 2 (Male)

Gender	Y	S
M1	1	0.8
M2	1	0.8
M3	1	0.7
M4	0	0.7
M5	0	0.3
M6	0	0.3

where the threshold for intervention (withdrawing a child from his family) depends on the available seats in foster houses.

Well-calibration [6] is a stronger variant of calibration. It requires that (1) calibration is satisfied, (2) the score is interpreted as the probability to truly belong to the positive class, and (3) for each score $S = s$, the probability to truly belong to the positive class is equal to that particular score:

$$P(Y = 1 \mid S = s, A = 0) = P(Y = 1 \mid S = s, A = 1) = s \quad \forall\, s \in [0, 1] \tag{15}$$

Intuitively, for a set of applicants who have a certain probability s of being hired, approximately s percent of these applicants should truly be hired. Table 11 (a) is a job hiring scenario which is calibrated (the proportion of applicants which should be hired for every score value is the same for male and female groups) but not well-calibrated (the score value does not coincide with the proportion of applicants that should be hired). Table 11 (b) is both calibrated and well-calibrated. Garg et al. [82] show that the difference between calibration and well-calibration is a simple difference in mapping. That is, "the scores of a calibrated predictor can, using a suitable transformation, be converted to scores satisfying well-calibration".

5.9. Group vs individual fairness notions

All the fairness notions discussed above are considered as group fairness where their common objective is to ensure that groups who differ by their sensitive attributes are treated equally. These

Table 11: 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

notions, mainly based on statistical measures, generally ignore all attributes of the individuals except the sensitive attribute A. Such treatment might hide unfairness. Dwork et al. [70] stated that group fairness, despite its suitability for policies among demographic sub-populations, does not guarantee that individuals are treated fairly. This is illustrated in the simple example in Table 12. The example satisfies most of group fairness notions, including total fairness (Section 5.6). However, based on the applicants profiles, it is clear that the predictor is unfair towards applicant Female 4. The fairness notions which follow attempt to address such issues by not marginalizing over non-sensitive attributes X of an individual, therefore they are called individual fairness notions [23].

Table 12: A simple job hiring example satisfying most of group fairness notions, but unfair towards Female 4 applicant.

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

TPR = 0.5
FPR = 0.66
PPV = 0.33
OA = 0.4

TPR = 0.5
FPR = 0.66
PPV = 0.33
OA = 0.4

[23] The term individual fairness is used in some papers to refer to fairness through awareness (Section 5.11). In this paper, the term individual fairness refers to fairness notions which cannot be considered as group fairness notions.

5.10. Causal discrimination

Causal Discrimination [83] implies that a classifier should produce exactly the same prediction for individuals who differ only from gender while possessing identical attributes X. In our hiring example, this is to say that male and female applicants with the same attributes X should have the same predictions:

$$X_{(\mathrm{A}=0)} = X_{(\mathrm{A}=1)} \ \wedge \ A_{(\mathrm{A}=0)} \ \neq A_{(\mathrm{A}=1)} \ \Rightarrow \hat{y}_{(\mathrm{A}=0)} = \hat{y}_{(\mathrm{A}=1)} \tag{16}$$

In our example, this implies that male and female applicants who otherwise have the same attributes X will either both be assigned a positive prediction or both assigned a negative prediction. Considering the example of Table 2, two applicants of different genders (Female 2 and Male 4) have identical values of X yet, getting different predictions (negative for female applicant while positive for male applicant). The predictor is then unfair towards Female 2 applicant.

At a first glance, causal discrimination can be seen as an extreme case of conditional statistical parity (Section 5.2) when conditioning on all non-sensitive attributes ($E = X$). However, conditional statistical parity is a group fairness notion which is satisfied if the proportion of individuals having the same non-sensitive attribute values and predicted accepted in both groups (e.g. male and female) is the same. This is why Eq. 2 is expressed in terms of conditional probabilities. Causal discrimination, however, consider every individual separately regardless of its contribution to sub-population proportions. To illustrate this subtlety, consider the following scenario:

Female 1	8	2	26	single	$\hat{Y} = 0$
Female 2	8	2	26	single	$\hat{Y} = 1$
Male 1	8	2	26	single	$\hat{Y} = 1$
Male 2	8	2	26	single	$\hat{Y} = 0$

Conditional statistical parity with $E = X$ (conditioning on all non-sensitive attributes) is satisfied as the proportion of males and females having the exact same attribute values and predicted accepted is the same (0.5). However, at the individual level, causal discrimination is not satisfied as there are two violations: Female 1 vs Male 1 and Female 2 vs Male 2. The two violations compensated each others and as a result conditional statistical parity is satisfied.

Causal discrimination is suitable to use in decision making scenarios where it is very common to find individuals sharing exactly the same attribute values. For example, admission decision making based mainly on test scores and categorical attributes. To apply this fairness notion on a loan granting scenario where there are only few individuals with exactly the same attribute values, Verma and Rubin [13] generated, for every applicant in the dataset, an identical individual of the opposite gender. The result of applying causal discrimination is the percentage of violations in the entire population (i.e. how many individuals are unfairly treated?).

5.11. Fairness through awareness

Fairness through awareness [70] (a.k.a individual fairness [4, 71]) is a generalization of causal discrimination which implies that similar individuals should have similar predictions. Let i and j be two individuals represented by their attributes values vectors v_i and v_j. Let $d(v_i, v_j)$ represent the similarity distance between individuals i and j. Let $M(v_i)$ represent the probability distribution over the outcomes of the prediction. For example, if the outcome is binary (0 or 1), $M(v_i)$ might be $[0.2, 0.8]$ which means that for individual i, $P(\hat{Y} = 0) = 0.2$ and $P(\hat{Y} = 1) = 0.8$. Let D be a distance metric between probability distributions. Fairness through awareness is achieved iff, for any pair of individuals i and j:

$$D(M(v_i), M(v_j)) \leq d(v_i, v_j) \tag{17}$$

For our hiring example, this implies that the distance between the distribution of outcomes of two applicants should be at most the distance between those applicants[24]. A possible relevant features to use for measuring the similarity between two applicants might be the education level and the job experience. The distance metric d between two applicants could be defined as the average of the normalized difference (the difference divided by the maximum difference in a dataset) of their education level and their job experience. More formally, let E_{v_i} and E_{v_j} be the education levels of individuals i and j, respectively, and let N_E be the normalized difference between education levels, that is, $N_E = \frac{|E_{v_i} - E_{v_j}|}{m_E}$ where m_E is the maximum difference in education level in the dataset. Similarly, let J_{v_i} and J_{v_j} be the job experience of individuals i and j, while N_J is the normalized

[24] Reducing all difference between two applicants/instances to a single distance value is often not easy to do in practice.

difference of the job experience, that is, $N_J = \frac{|J_{v_i} - J_{v_j}|}{m_J}$ where m_J is the maximum difference in job experience in the dataset. The distance metric is defined as:

$$d(v_i, v_j) = \frac{N_E + N_J}{2},$$

The distance between the probability distributions over the outcomes could be the *Hellinger distance* [84]. Let $\{y_1, y_2, \ldots, y_K\}$ be the set of possible outcomes and let P and Q two (discrete) probability distributions. The Hellinger distance between P and Q is defined as:

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^{K} \left(\sqrt{P(y_k)} - \sqrt{Q(y_k)} \right)^2}$$

Table 13 shows a sample from the job hiring dataset on which fairness through awareness is applied. The result of applying fairness through awareness is shown in Table 14. Each cell at the left of the shaded diagonal represents a distance between two individuals and each cell at the right of the shaded diagonal represents the distance between probability outcomes of two individuals.
For instance:

$$d(F1, F2) = 0.25$$

While:

$$\begin{aligned} D(M(F1), M(F2)) &= \frac{1}{\sqrt{2}} \sqrt{\left(\sqrt{0.4} - \sqrt{0.3} \right)^2 + \left(\sqrt{0.6} - \sqrt{0.7} \right)^2} \\ &= \frac{1}{\sqrt{2}} \sqrt{0.0081 + 0.0036} \\ &= 0.07 \end{aligned}$$

The cell values in bold represent the cases where fairness through awareness is not satisfied: $D \not\leq d$. For example, **0.07** (< 0.0) implies that F_1 is discriminated compared to M_3. Similarly, M_2 is discriminated compared to F_3, F_2, and M_3.

Fairness through awareness is more fine-grained than any group fairness notion presented earlier in Sections 5.1– 5.8. For instance, in the example of Table 13, statistical parity is satisfied: 0.33 for both men and women. Likewise, equalized odds (5.3) is satisfied as the TPR and the FPR are equal for male and female applicants (0.5 and 0, respectively). Nevertheless, Table 14 shows that

Table 13: Job hiring sample used to apply fairness through awareness.

(a) Dataset						(b) Prediction	
Gender	Education Level	Job Experience	Age	Marital Status	Label	$\hat{Y}$	S
Female 1	12	2	39	single	1	0	0.4
Female 2	12	1	26	married	0	0	0.3
Female 3	13	1	32	married	1	1	0.9
Male 1	13	1	26	married	1	0	0.2
Male 2	12	1	41	single	0	0	0.2
Male 3	12	2	30	single	1	1	0.7

Table 14: Application of fairness through awareness. Each cell at the left of the shaded table's diagonal represents a distance between a pair of applicants. Those at the right represent the distance between probability distributions. Values in bold imply cases where $D > d$, meaning fairness through awareness is not satisfied.

	F1	F2	F3	M1	M2	M3
F1		0.07	0.26	0.16	0.16	**0.07**
F2	0.25		0.18	0.08	**0.08**	**0.29**
F3	0.75	0.5		0.1	**0.54**	0.18
M1	0.75	0.5	0.0		0.0	0.08
M2	0.25	0.0	0.5	0.5		**0.37**
M3	0.0	0.25	0.75	0.75	0.25	

$d(v_i, v_j)$

$D(M(v_i), M(v_j))$

when comparing each pair of individuals (regardless of their gender) cases of discrimination have been discovered.

It is important to mention that, in practice, fairness through awareness introduces some challenges. For instance, it assumes that the similarity metric is known for each pair of individuals [85]. That is, a challenging aspect of this approach is the difficulty to determine what is an appropriate metric function to measure the similarity between two individuals. Typically, this requires careful human intervention from professionals with domain expertise [71]. For instance, suppose a company is intending to hire only two employees while three applicants i_1, i_2 and i_3 are eligible for the offered job. Assume i_1 has a bachelor's degree and 1 year related work experience, i_2 has a master's degree and 1 year related work experience and i_3 has a master's degree but no related work experience (Figure 2). Is i_1 closer to i_2 than i_3? If so, by how much? This is difficult to answer, especially if the company overlooked such specific cases and did not carefully define and set a suitable and fair

similarity metric in order to rank applicants for job selection. Thus, fairness through awareness can not be considered suitable for domains where trustworthy and fair distance metric is not available.

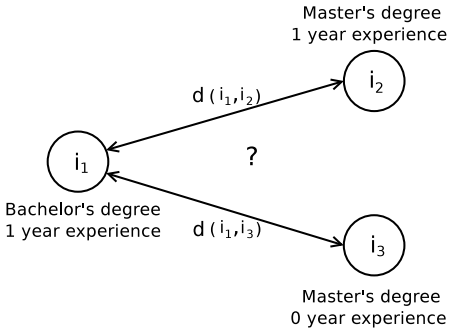

Figure 2:
An example showing the difficulty of selecting a distance metric in fairness through awareness

5.12. Causality-based fairness notions

Causality-based fairness notions differ from all aforementioned statistical fairness approaches in that they are not totally based on data but consider additional knowledge about the structure of the world, in the form of a causal model. This additional knowledge helps us understand how data is generated in the first place and how changes in variables propagate in a system. 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 [86]. Causal relationships are expressed using structural equations [87] and represented by causal graphs where nodes represent variables (attributes) and edges represent causal relationships between variables. Figure 3 shows a possible causal graph for our hiring example where directed edges indicate causal relationships.

Statistical parity (Section 5.1) is known also as total variation (TV) as it can be expressed by subtracting the two terms in Eq. 1 as follows:

$$TV_{a_1,a_0}(\hat{y}) = P(\hat{Y} = \hat{y} \mid A = a_1) - P(\hat{Y} = \hat{y} \mid A = a_0) \tag{18}$$

A TV equal zero indicates fairness according to statistical parity. As TV is purely a statistical

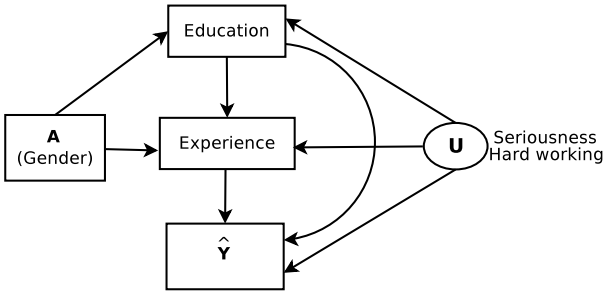

Figure 3: A possible causal graph for the hiring example.

notion, it is unable to reflect the causal relationship between A and Y, that is, it is insensitive to the mechanism by which data is generated.

Total effect (TE) [88] is the causal version of TV and is defined in terms of experimental probabilities as follows :

$$TE_{a_1,a_0}(\hat{y}) = P(\hat{y}_{A\leftarrow a_1}) - P(\hat{y}_{A\leftarrow a_0}) \tag{19}$$

$P(\hat{y}_{A\leftarrow a}) = P(\hat{Y} = \hat{y} \mid do(A = a))$ is called an experimental probability and is expressed using intervention. An intervention, noted $do(V = v)$, 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 $P(\hat{Y} = 1 \mid A = 0)$ reflects the probability of hiring among female applicants, $P(\hat{Y}_{A\leftarrow 0} = 1 = P(\hat{Y} = 1) \mid do(A = 0))$ reflects the probability of hiring if *all the candidates in the population* had been female. The obtained distribution $P(\hat{Y}_{A\leftarrow a})$ 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 noted also $\hat{Y}_{A=a}$ 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 TV 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 attribute value (e.g. gender) on the outcome (e.g. probability of hiring) given that we already observed the outcome for that individual. This typically involves an impossible situation which requires to go back in the past and change the sensitive attribute value. Mathematically, this can be formalized using counterfactual quantities. The simplest fairness notion using counterfactuals is the effect of treatment on the

treated (ETT) [88].

The effect of treatment on the treated (ETT) is defined as:

$$ETT_{a_1,a_0}(\hat{y}) = P(\hat{y}_{A\leftarrow a_1} \mid a_0) - P(\hat{y} \mid a_0) \tag{20}$$

$P(\hat{y}_{A\leftarrow a_1} \mid a_0)$ 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, $P(\hat{Y}_{A\leftarrow 1} \mid A = 0)$ reads the probability of hiring an applicant had she been a male, given that the candidate is observed to be female. Such probability involves two worlds: an actual world where $A = a_0$ (the candidate is female) and a counterfactual world where for the same individual $A = a_1$ (the same candidate is male). Notice that $P(\hat{y}_{A\leftarrow a_0} \mid a_0) = P(\hat{y} \mid a_0)$, a property called consistency [88].

Counterfactual fairness [71] is a fine-grained variant of ETT conditioned on all attributes. That is, a prediction $\hat{Y}$ is counterfactually fair if under any assignment of values $X = x$,

$$P(\hat{Y}_{A\leftarrow a_1} = \hat{y} \mid X = x, A = a_0) = P(\hat{Y}_{A\leftarrow a_0} = \hat{y} \mid X = x, A = a_0) \tag{21}$$

where X is the set of all attributes excluding A. Since conditioning is done on all remaining variables X, counterfactual fairness is an individual notion. According to Eq. 21, counterfactual fairness is satisfied if the probability distribution of the outcome $\hat{Y}$ is the same in the actual and counterfactual worlds, for every possible individual. In the job hiring example, an MLDM is counterfactually fair if:

$$P(\hat{Y}_{A\leftarrow 1} \mid X = x, A = 0) = P(\hat{Y}_{A\leftarrow 0} \mid X = x, A = 0) \tag{22}$$

The main problem with the applicability of TE, ETT, and counterfactual fairness is the computation of the non-observable terms in Eqs 19, 20, and 21. These terms are either interventional (e.g. $P(\hat{y}_{A\leftarrow a_1})$) or counterfactual (e.g. $P(\hat{Y}_{A\leftarrow a_1} = \hat{y} \mid X = x, A = a_0)$. In scenarios where these quantities can be expressed in terms of observable probabilities (e.g. joint probabilities, conditional probabilities, etc.), it is said that they are *identifiable*. Otherwise, they are unidentifiable. Typically, the identifiability of interventional and counterfactual quantities depends on the structure of the causal graph [89, 88]. Alternatively, if all parameters of the causal model are known (including the latent variables distributions $P(U = u)$), any counterfactual is identifiable and can be computed using the three steps abduction, action, and prediction (Theorem 7.1.7 in [88]). The details of

the computation of a counterfactual probability using a simple deterministic example are provided in Appendix A.

A simple but important implication of Eq. 21 is that, given a causal graph, a predictor $\hat{Y}$ is counterfactually fair if it is a function of non-descendants of the sensitive variable A. In other words, if $\hat{Y}$ is a function of variables that depend on A (there is a directed path between any one of those variables and A), it is not counterfactually fair. Consequently, one can tell if a predictor is counterfactually fair by simply checking the causal graph[25].

No unresolved discrimination [90] is another causal-based fairness notion which is satisfied when no directed paths from the sensitive attribute A to the predictor $\hat{Y}$ are allowed, except via a resolving variable. A resolving variable is any variable in a causal graph that is influenced by the sensitive attribute in a manner that is accepted as nondiscriminatory (this is similar to explanatory attributes in conditional statistical parity (Section 5.2)). In the job hiring example, if we assume that the effect of A on the education level is nondiscriminatory, it implies that the differences in education level for different values of A are not considered as discrimination. Thus, a disparity in the predictions between men and women might been explained and justified by their corresponding education levels. Hence, the education level acts as a resolving variable. Figure 4 shows two similar causal graphs for our hiring example, yet differ in some of the causal relations between variables. By considering the education as a resolving variable, the graph at the left exhibits unresolved discrimination along the dashed paths: $A \rightarrow Experience \rightarrow \hat{Y}$ and $A \rightarrow \hat{Y}$. By contrast, the graph at the right does not exhibit any unresolved discrimination as the effect of A on $\hat{Y}$ is justified by the resolved variable Education: $A \rightarrow Education \rightarrow \hat{Y}$.

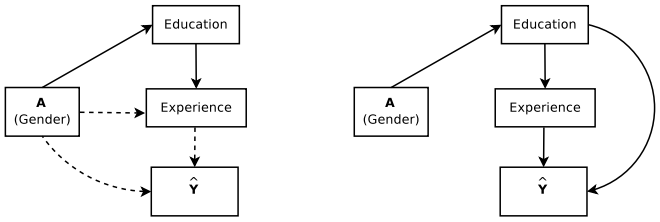

Figure 4: Two possible graphs for the hiring example. If *Education* is a resolving variable, the predictor $\hat{Y}$ exhibits unresolved discrimination in the left graph (along the dashed paths), but not in the right one.

[25]Kusner et al. [71] identify some exceptions, but guaranteeing that they will *not happen in general*.

No unresolved discrimination is equivalent to other fairness notions in some interesting special cases [90]. For instance, if no resolving variables exist, no unresolved discrimination is analogous to statistical parity (Section 5.1) in a causal context. A and $\hat{Y}$ are statistically independent and no directed paths from A to $\hat{Y}$ are allowed. Likewise, no unresolved discrimination might be equivalent to equalized odds (Section 5.3) in a causal context if the set of resolving variables is the singleton set of actual outcomes: $\{Y\}$. Compared to counterfactual fairness, no unresolved discrimination is a weaker notion. That is, a counterfactually unfair scenario may be identified as fair based on no unresolved discrimination. This can happen in case one or several variables in the causal graph are identified as resolving.

A causal graph exhibits potential proxy discrimination [90] if there exists a path from the protected attribute A to the predicted outcome $\hat{Y}$ that is blocked by a proxy variable P_x. A proxy is a descendant of A that is chosen to be labelled as a proxy because it is significantly correlated with A. Given a causal graph, a predictor $\hat{Y}$ exhibits no proxy discrimination if following equality holds for all potential proxies P_x.

$$P(\hat{Y}_{P_x \leftarrow p}) = P(\hat{Y}_{P_x \leftarrow p\prime}) \quad \forall\, p, p\prime \tag{23}$$

In other words, Eq. 23 implies that changing the value of P_x should not have any impact on the prediction. In the job hiring example, the job experience can be considered as a proxy of an individual's gender. Figure 5 shows two similar causal graphs. The one at the left presents a potential proxy discrimination via the path: $A \rightarrow \textit{Experience} \rightarrow \hat{Y}$. However, the graph at the right is free of proxy discrimination as the edge between A and its proxy P_x (here Experience) has been removed along with all incoming arrows of P_x (the edge between *Education* and *Experience*).

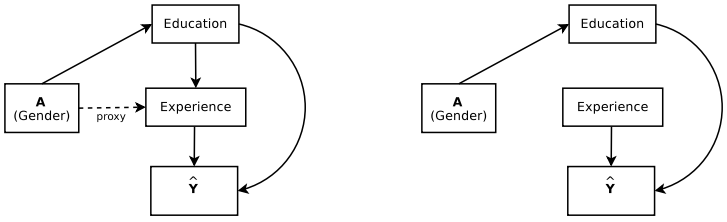

Figure 5: Two possible graphs to describe proxy discrimination. If we consider *Experience* as a proxy of the sensitive attribute A, the graph at the left exhibits a potential proxy discrimination (along the dashed edge between A and Experience), but not in the right one.

Other causal based fairness notions include direct/indirect effect [91], FACE/FACT [92], counterfactual effects [93], counterfactual error rates [94], and path-specific counterfactual fairness [95, 96].

As a general rule, causality-based fairness notions can be used as long as the causal relationships between the attributes are identified and represented using a reliable and plausible causal graph. The construction of the causal graph requires typically domain-specific expertise and can be validated by existing datasets. In practice, however, causality-based fairness notions are recommended in at least two notable scenarios. The first scenario is when the legal framework of the case at hand is disparate treatment. In such framework, to win a discrimination case, the plaintiff must show that the defendant has used (directly or indirectly (via proxy)) the sensitive attribute A to take the discriminatory decision $\hat{Y}$. In other words, she must prove that the variable A is a cause of $\hat{Y}$ while the causal effect of A on $\hat{Y}$ is central to all causal-based fairness notions mentioned above. The second scenario is when there is confounding between A and $\hat{Y}$. That is, there is a covariate which is a common cause of A and $\hat{Y}$. Such scenario can lead to statistical anomalies such as Simpson's paradox [97, 88] where the statistical conclusions drawn from the sub-populations differ from that from the whole population. The Berkeley admission case [98] is a known real-world example of such statistical anomaly. In such scenarios, any statistical fairness notion which relies solely on correlation between variables, will fail to detect bias. Hence, causality-based fairness notions are necessary to appropriately address the problem of fairness.

6. Relaxation

Almost all fairness notions presented so far involve a strict equality between quantities, in particular probabilities. In real scenarios, however, it is more suitable to opt for an approximate or relaxed form of fairness constraint. The need for relaxation might be due to the impossibility to apply fairness strictly on the application at hand, or merely, it is not a requirement to impose an exact constraint [99].

Fairness notion definitions can be relaxed by considering a threshold on the ratio or difference between quantities. For instance, the requirement for statistical parity (Section 5.1) can be relaxed in one of the two following ways:

- By allowing the ratio between the predicted acceptance rates of protected and unprotected groups to reach the threshold of ϵ (a.k.a $p\%$ rule defined as satisfying this inequality when

$\epsilon = p/100$ [100]):

$$\frac{P(\hat{Y} \mid A = 0)}{P(\hat{Y} \mid A = 1)} \geq 1 - \epsilon \quad \forall\, \epsilon \in [0, 1] \tag{24}$$

For $\epsilon = 0.2$, this condition relates to the 80% rule in disparate impact law [101, 64].

- By allowing the difference between the predicted acceptance rates of different groups to reach a threshold of ϵ [70]:

$$\mid P(\hat{Y} \mid A = 0) - P(\hat{Y} \mid A = 1) \mid \leq \epsilon \quad \forall\, \epsilon \in [0, 1] \tag{25}$$

A notable difference between the two types of relaxation is that the second one (Eq. 25) is insensitive to which group/individual is the victim of discrimination as the formula is using absolute value.

Fairness through awareness can be relaxed using three threshold values, α_1, α_2, and γ as follows [102]:

$$P\Big[P[|M(v_i) - M(v_j)| > d(v_i, v_j) + \gamma\ ] > \alpha_2\Big] \leq \alpha_1. \tag{26}$$

The relaxation is allowing $M(v_i) - M(v_j)$ to exceed $d(v_i, v_j)$ by a margin of γ, but the fraction of individuals differing from them by γ should not exceed α_2. If the fraction exceeds α_2, the individual is said to be α_2-discriminated against.

To allow for more flexibility in the application of fairness notions, other relaxations can be considered. For instance, Eq. 2 of conditional statistical parity (Section 5.2) can be modified by relaxing the strict equality $E = e$ as follows:

$$P(\hat{Y} = 1 \mid e - \epsilon \leq E \leq e + \epsilon, A = 0) = P(\hat{Y} = 1 \mid e - \epsilon \leq E \leq e + \epsilon, A = 1) \tag{27}$$

7. Classification and tensions

Group fairness notions fall into three classes defined in terms of the properties of joint distributions, namely, independence, separation, and sufficiency [8]. These properties are used in the literature to prove the existing of tensions between fairness notions, that is, it is impossible to satisfy all

fairness notions simultaneously except in extreme, degenerate, and dump scenarios. Besides, the applicability of most of fairness notions can be ameliorated by relaxing their strict definitions.

7.1. Classification

Group fairness (a.k.a statistical fairness) notions can be characterized by the properties of the joint distribution of the sensitive attribute A, the label Y, and the classifier $\hat{Y}$ (or score S). This means that we can write them as some statement involving properties of these three random variables resulting in the three following fairness criteria [72, 8]:

Independence. Independence means that the sensitive feature A is statistically independent of the classifier $\hat{Y}$ (or the score S).

$$\hat{Y} \perp A \quad (or\ S \perp A) \tag{28}$$

In the case of binary classification, independence is equivalent to statistical parity as defined in Section 5.1, Eq. 1. Conditioning on explanatory variables (E) yields a variant of independence as follows.

Conditional independence.

$$\hat{Y} \perp A \mid E \quad (or\ S \perp A \mid E) \tag{29}$$

This class includes conditional statistical parity defined in Section 5.2, Eq. 2.

Separation. Separation denotes a class of fairness notions satisfying, at different degrees, conditional independence between the prediction $\hat{Y}$ and the sensitive attribute A given the actual outcome Y.

$$\hat{Y} \perp A \mid Y \quad (or\ S \perp A \mid Y) \tag{30}$$

In the case where $\hat{Y}$ is a binary classifier, the formulation of separation is equivalent to that of the equalized odds (Eq. 3). Equal opportunity (Eq. 4), predictive equality (Eq. 5), balance for positive class (Eq. 11), and balance for negative class (Eq. 12) are all relaxations of separation. Some incompatibility results do hold for separation, but do not hold for the relaxations. More on this in the next section (Section 7.2).

Sufficiency. Sufficiency is a class of fairness notions satisfying, at different degrees, conditional independence between the target variable Y and the sensitive attribute A given the prediction $\hat{Y}$.

$$Y \perp A \mid \hat{Y} \quad (or\ Y \perp A \mid S) \tag{31}$$

In the case of binary classification, strict sufficieny corresponds to conditional use accuracy equality (Eq. 6). Using the score S, calibration (Eq. 14), and well-calibration (Eq. 15) can be considered as sufficiency [9]. Relaxation of sufficiency yields to predictive parity (Eq. 7) which also does not satisfy exactly the same incompatibility result as sufficiency (Section 7.2).

Table 15 lists all fairness notions along with their classification.

Table 15: Classification of fairness notions. ($*$ notion newly defined in this paper)

Fairness Notion	Ref.	Formulation	Classification	Type
Statistical parity	[70]	$P(\hat{Y} \mid A=0) = P(\hat{Y} \mid A=1)$	Independence (equivalent or relaxed$^\bigstar$)	Group
Conditional statistical parity	[7]	$P(\hat{Y}=1 \mid E=e, A=0) = P(\hat{Y}=1 \mid E=e, A=1)^\bigstar$		
Equalized odds	[79]	$P(\hat{Y}=1 \mid Y=y,\ A=0) = P(\hat{Y}=1 \mid Y=y,\ A=1) \quad \forall y \in \{0,1\}$	Separation (equivalent or relaxed$^\bigstar$)	
Equal opportunity		$P(\hat{Y}=1 \mid Y=1, A=0) = P(\hat{Y}=1 \mid Y=1, A=1)^\bigstar$		
Predictive equality	[7]	$P(\hat{Y}=1 \mid Y=0, A=0) = P(\hat{Y}=1 \mid Y=0, A=1)^\bigstar$		
Balance for positive class	[6]	$E[S \mid Y=1, A=0)] = E[S \mid Y=1, A=1]^\bigstar$		
Balance for negative class		$E[S \mid Y=0, A=0] = E[S \mid Y=0, A=1]^\bigstar$		
Overall balance	*	$E[S \mid Y=y, A=0] = E[S \mid Y=y, A=1] \quad \forall y \in \{0,1\}$		
Conditional use acc. equality	[12]	$P(Y=y \mid \hat{Y}=y, A=0) = P(Y=y \mid \hat{Y}=y, A=1) \quad \forall y \in \{0,1\}$	Sufficiency (equivalent or relaxed$^\bigstar$)	
Predictive parity	[9]	$P(Y=1 \mid \hat{Y}=1, A=0) = P(Y=1 \mid \hat{Y}=1, A=1)^\bigstar$		
Negative predictive parity	*	$P(Y=1 \mid \hat{Y}=0, A=0) = P(Y=1 \mid \hat{Y}=0, A=1)^\bigstar$		
Calibration	[9]	$P(Y=1 \mid S=s, A=0) = P(Y=1 \mid S=s, A=1) \quad \forall s \in [0,1]$		
Well-calibration	[6]	$P(Y=1 \mid S=s, A=0) = P(Y=1 \mid S=s, A=1) = s \quad \forall\, s \in [0,1]$		
Overall accuracy equality	[12]	$P(\hat{Y}=Y \mid A=0) = P(\hat{Y}=Y \mid A=1)$	Other metrics from confusion matrix	
Treatment equality		$\frac{FN}{FP}_{(\mathrm{A}=0)} = \frac{FN}{FP}_{(\mathrm{A}=1)}$		
Total fairness		$-$	Independence, Separation and Sufficiency	
Total effect	[88]	$TE_{a_1,a_0}(\hat{y}) = P(\hat{y}_{A \leftarrow a_1}) - P(\hat{y}_{A \leftarrow a_0})$	Causality	
Effect of treatment on treated		$ETT_{a_1,a_0}(\hat{y}) = P(\hat{y}_{A \leftarrow a_1} \mid a_0) - P(\hat{y} \mid a_0)$		
No unresolved discrimination	[90]	$-$		
No proxy discrimination		$P(\hat{Y} \mid do(P_x = p)) = P(\hat{Y} \mid do(P_x = p\prime)) \quad \forall P_x \quad and \quad \forall\, p, p\prime$		
Counterfactual fairness	[71]	$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X=x, A=a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X=x, A=a)$		Individual
Causal discrimination	[83]	$X_{(\mathrm{A}=0)} = X_{(\mathrm{A}=1)} \ \wedge\ A_{(\mathrm{A}=0)} \neq A_{(\mathrm{A}=1)} \ \Rightarrow \hat{y}_{(\mathrm{A}=0)} = \hat{y}_{(\mathrm{A}=1)}$	Similarity Metric	
Fairness through awareness	[70]	$D(M(v_i), M(v_j)) \leq d(v_i, v_j)$		

7.2. Tensions

It has been proved that there are incompatibilities between fairness notions. That is, it is not always possible for an MLDM to satisfy specific fairness notions simultaneously [72, 8, 9, 5, 3]. In presence of such incompatibilities, the MLDM should make a trade-off to satisfy some notions on the expense of others or partially satisfy all of them. Incompatibility[26] results are well summarized by Mitchell et al. [3] as follows:

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

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

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 an MLDM to satisfy statistical parity and predictive parity between two groups having different base rates. Such example needs to satisfy the following constraints, assuming two groups a and b:

$$\frac{TP_a+FP_a}{TP_a+FP_a+FN_a+TN_a} = \frac{TP_b+FP_b}{TP_b+FP_b+FN_b+TN_b} \quad \text{(independence)}$$

$$\frac{TP_a}{TP_a+FP_a} = \frac{TP_b}{TP_b+FP_b} \quad \text{(predictive parity)}$$

$$\frac{TP_a+FN_a}{TP_a+FP_a+FN_a+TN_a} \neq \frac{TP_b+FN_b}{TP_b+FP_b+FN_b+TN_b} \quad \text{(different base rates)}$$

An example scenario satisfying the above constrains is the following:

$PPV_a = 0.4$	$TP_a = 9$	$FP_a = 6$	$TP_b = 12$	$FP_b = 8$	$PPV_b = 0.4$
$baserate_a = 0.43$	$FN_a = 4$	$TN_a = 11$	$FN_b = 2$	$TN_b = 18$	$baserate_b = 0.35$

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

Statistical parity (independence) versus equalized odds (separation). Similar to the previous result, independence and separation are mutually exclusive unless base rates are equal or the predictor $\hat{Y}$ is independent from the actual label Y [8]. As mentioned earlier, dependence between $\hat{Y}$ and Y is a weak assumption as any useful predictor should satisfy it. More formally,

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

Considering a relaxation of equalized odds, that is, equal opportunity or predictive equality, breaks the incompatibility between independence and separation. An MLDM whose output satisfies independence and equal opportunity, but with different base rates between groups should satisfy the following constraints:

$$\frac{TP_a+FP_a}{TP_a+FP_a+FN_a+TN_a} = \frac{TP_b+FP_b}{TP_b+FP_b+FN_b+TN_b} \quad \text{(independence)}$$

$$\frac{TP_a}{TP_a+FN_a} = \frac{TP_b}{TP_b+FN_b} \quad \text{(equal opportunity)}$$

$$\frac{TP_a+FN_a}{TP_a+FP_a+FN_a+TN_a} \neq \frac{TP_b+FN_b}{TP_b+FP_b+FN_b+TN_b} \quad \text{(different base rates)}$$

An example scenario satisfying the above constrains is the following:

$TPR_a = 0.6$	$TP_a = 9$	$FP_a = 3$	$TP_b = 12$	$FP_b = 6$	$TPR_b = 0.6$
$baserate_a = 0.55$	$FN_a = 2$	$TN_a = 6$	$FN_b = 8$	$TN_b = 4$	$baserate_b = 0.71$

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

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

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 [9]:

- 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. For example, the following scenario satisfies equal opportunity (relaxed version of separation) and predictive parity (relaxed version of sufficiency) while base rates are different in both groups:

$TPR_a = 0.4$					$TPR_b = 0.4$
	$TP_a = 9$	$FP_a = 6$		$TP_b = 12$	$FP_b = 8$	
$PPV_a = 0.75$					$PPV_b = 0.75$
	$FN_a = 3$	$TN_a = 2$		$FN_b = 4$	$TN_b = 8$	
$baserate_a = 0.6$					$baserate_b = 0.5$

7.3. Group vs individual fairness

Compared to individual fairness notions, the main concern for group fairness notions is that they are only suited to a limited number of coarse-grained, predetermined protected groups based on some sensitive attribute (e.g. gender, race, etc.). Hence group fairness notions are not suitable in presence of intersectionality [63] where individuals are often disadvantaged by multiple sources of discrimination: their race, class, gender, religion, and other inner traits. Typically, statistical fairness can only be applied across a small number of coarsely defined groups, and hence failing to identify discrimination on structured subgroups (e.g. single women) known also as "fairness gerrymandering" [103]. A simple alternative might be to apply statistical fairness across every possible combination of protected attributes. There are at least two problems to this approach. First, this can lead to an impossible statistical problem with the large number of sub-groups which may lead in turn to overfitting. Second, groups which are not (yet) defined in anti-discrimination law may exist and may need protection [104]. Another issue with group fairness notions is their susceptibility to masking. Most of group fairness notions can be gamed by adding arbitrarily selected samples to satisfy the fairness notion formula, that is, to just "make up the numbers".

Compared to group fairness notions, individual fairness notions have the drawback that they can result in "unjust disparities in outcomes between groups" [105]. For illustration, consider the example in Table 16 where fairness through awareness is satisfied (Eq. 17) whereas statistical parity Eq. (1) is

not. Fairness through awareness is satisfied since for every pair of candidates, the distance between the probability distributions on the outcomes ($M()$) is smaller than the distance between the pair of candidates. On the other hand, if the hiring threshold is 0.6, only one female candidate ($F2$) will be hired as she has a probability of acceptance $P(\hat{Y} = 1) = 0.8 > 0.6$ whereas all male candidates will be hired. Another important issue for similarity-based individual fairness (e.g. fairness through awareness) is the difficulty to obtain a similarity value between every pair of individuals. For example, even with the assumption that the similarity can be quantified between all individuals in the training data, it might be challenging to generalize to new individuals [105].

Table 16: A job hiring scenario satisfying fairness through awareness (Eq. 17) but not statistical parity (Eq. 1) for a threshold of 0.6. The second row ($M()$) indicates the probability distribution on the outcomes. For example, for the first female applicant $F1$, $P(\hat{Y} = 1) = 0.58$ and $P(\hat{Y} = 0) = 0.42$. Each cell at the left of the shaded table's diagonal represents a distance between a pair of applicants. Those at the right represent the distance between probability distributions on the outcomes.

	F1	F2	F3	M1	M2	M3	
M()	$[0.58, 0.42]$	$[0.8, 0.2]$	$[0.55, 0.45]$	$[0.65, 0.35]$	$[0.81, 0.19]$	$[0.61, 0.39]$	
F1		0.17	0.021	0.051	0.18	0.02	$D(M(v_i), M(v_j))$
F2	0.21		0.19	0.11	0.008	0.15	
F3	0.06	0.22		0.07	0.20	0.04	
M1	0.1	0.15	0.1		0.12	0.029	
M2	0.2	0.01	0.3	0.15		0.15	
M3	0.05	0.17	0.08	0.05	0.17		
			$d(v_i, v_j)$				

Several researchers assume that both group and individual fairness are prominent, yet, conflicting and suggest approaches to minimize the trade-offs between these notions [105]. For instance, [10] define two different worldviews, 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 innate differences between groups of individuals based on certain potentially discriminatory characteristics. These two worldviews highlight the tension between group and individual fairness. For instance, in the job hiring example, the WYSIWYG might be the assumption that attributes like education level and job experience (which belong to the observed space) correlate well with the applicant's seriousness or hardworking (properties of the construct space). This is to say that there is some way to combine these two spaces to correctly compare true applicant aptitude for the job. On the other hand, the WAE claims that all groups will have almost the same distribution in the construct space of inherent

abilities (here, seriousness and hardworking), chosen as important inputs to the decision making process. The idea is that any difference in the groups' performance (e.g., academic achievement or education level) is due to factors outside their individual control (e.g., the quality of their neighborhood school) and should not be taken into account in the decision making process. Thus, the choice between fairness notions must be based on an explicit choice in worldviews.

8. Diagram and discussion

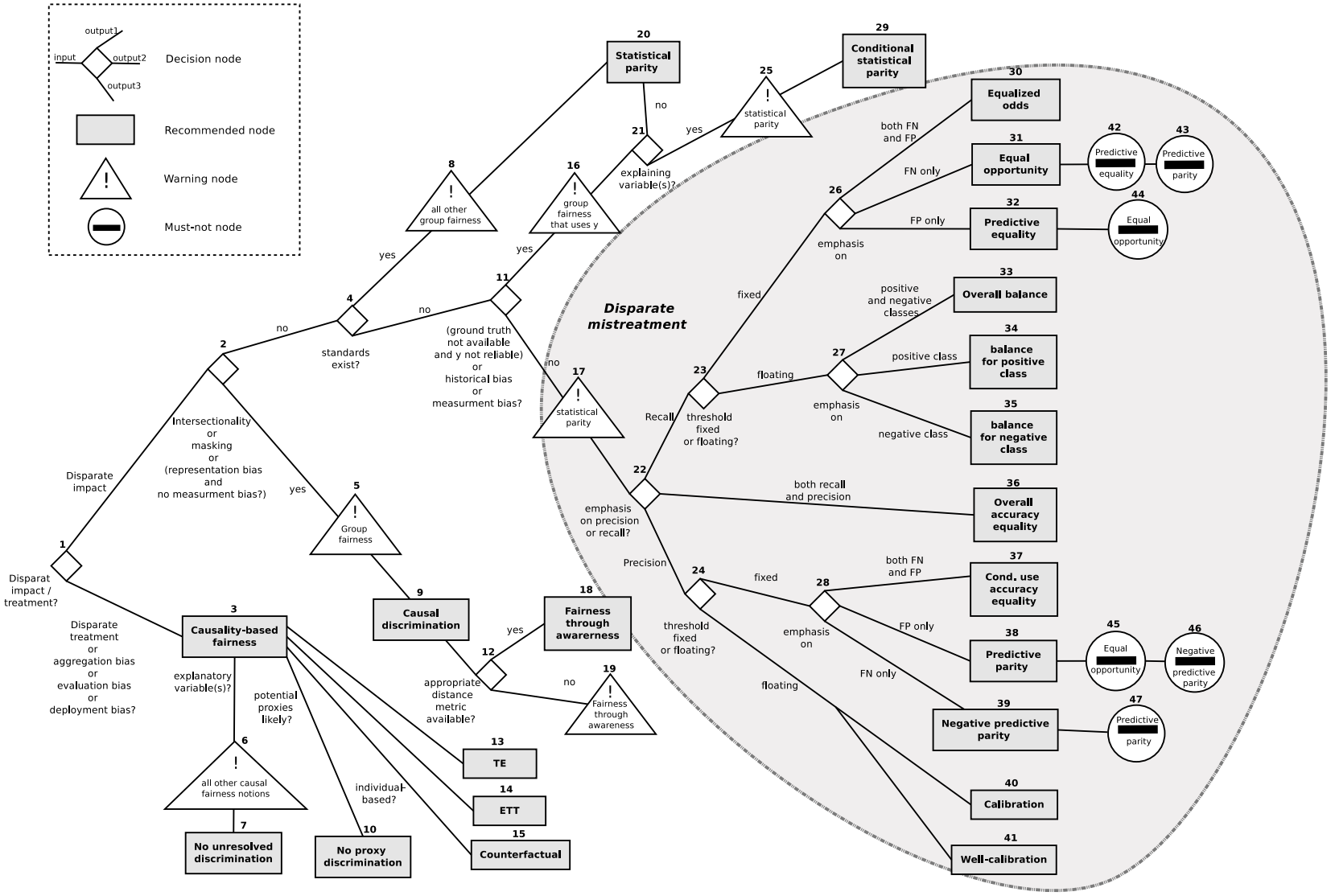

Figure 6: Fairness notions applicability decision diagram.

With the large number of fairness notions and the subtle resemblance between MLDM scenarios, deciding about which fairness notion to use is not a trivial task. More importantly, selecting and using a fairness notion in a scenario inappropriately may detect unfairness in an otherwise fair scenario, or the opposite, i.e., fail to identify unfairness in an unfair scenario.

One of the objectives of this survey is to systemize the selection procedure of fairness notions. This is achieved by identifying a set of fairness-related characteristics (Section 4) of the scenario at hand and then use them to recommend the most suitable fairness notion for that specific scenario. The proposed systemized selection procedure is illustrated in the decision diagram of Figure 6. The diagram is called "decision diagram" and not "decision tree" for the following reason. In typical decision trees, every leaf corresponds to a single decision, which is a fairness notion that *should* be used. However, the diagram in Figure 6 is designed such that every node indicates which notions are recommended, which notions to be avoided, and which notions must not be used. In addition, if a notion is not mentioned along the path, it means, it can be safely used.

The diagram is composed of four types of nodes:

- **Decision node (diamond):** based on fairness-related characteristics (Section 4).
- **Recommended node (rectangle):** a leaf node indicating that the fairness notion is suitable to be used given all fairness-related characteristics in the path to that node.
- **Warning node (triangle):** indicates that the fairness notion(s) is/are not recommended in all the branch in the right of the node. This node can appear in the middle of the edge between two decision nodes.
- **Must-not node (circle):** the fairness notion must not be used.

To illustrate how the diagram should be interpreted, consider the recommended node predictive parity (node 38). According to the diagram, predictive parity is recommended in the scenario where the legal framework is disparate impact (decision node 1), intersectionality and/or masking are unlikely (decision node 2), there is no evidence that representation bias is likely (decision node 2), standards do not exist (decision node 4), ground-truth is available or outcome Y is reliable (decision node 11), historical and measurement bias are unlikely (decision node 11), fairness is more sensitive to precision rather than recall (decision node 22), the prediction threshold is typically fixed (decision node 24) and the emphasis is on false positives rather than false negatives (decision node 28). In

that particular scenario, equal opportunity must not be used (must-not node 45) because fairness in this scenario is particularly sensitive to false positives, while equal opportunity is completely insensitive to false positives. Similarly, negative predictive parity must not be used (must-not node 46) as fairness is sensitive to precision rather than recall. The warning node 17 along the same path indicates that statistical parity is not suitable in this scenario. Finally, any fairness notion for which there is no a warning node or a must-not node along the path of the scenario can be used in this scenario. For instance, all individual fairness notions can be used.

As concrete example of situations where predictive parity (node 38) is recommended, consider the following. In situations when the outcome is influenced by the decision, some statistical quantities (e.g. FN, TN, etc.) are unlikely to be observed, and hence, any fairness notion that is defined in terms of those quantities is not suitable to use. For example, in real-world cases of loan-granting, a loan application which is predicted to be defaulting, will not be approved. Consequently, both negative statistics (true negative (TN) and false negative (FN)) will not be typically observed. Hence, fairness notions such as equalized odds and equality of opportunity cannot be used as they are defined in terms of TN and FN. In such cases, predictive parity (node 38) is recommended.

Node 1: Assessing fairness is very often performed in the context of a legal case where a plaintiff is filing a claim against a party that is using an MLDM. According to real-world legislation, in particular, the American anti-discrimination law, this can fall into one the two legal frameworks, namely, disparate impact and disparate treatment. If the plaintiff is filing the claim under the disparate impact framework, she can prove the liability of the defendant by using an observational group or individual fairness notion as the goal is to show that the practices and policies used by the defendant are facially neutral but have a disproportionately adverse impact on the protected class [64]. If, however, the plaintiff is filing a claim under the disparate treatment framework, observational fairness notions are often not enough to prove the liability of the defendant as the goal is to show that the defendant has used the sensitive attribute to take the discriminatory decision. The recommended fairness notions in that case are causality-based (recommended node 3) since all of them are expressed in terms of the causal effect of the sensitive attribute on the prediction.

Node 2: As explained above, any unintentional type of bias can also be "orchestrated" intentionally by decision makers with prejudicial views. For instance, decision makers can purposefully bias the data collection step to ensure that the MLDM remains less favorable to protected classes. To reliably assess the bias in presence of such masking attempts, all group fairness notions should

be avoided as they are defined in terms of statistics about the different sub-populations and hence can more easily be gamed by prejudicial decision makers. Intersectionality is similar to masking as both lead to a discrimination which is difficult to detect using statistical measures and consequently requires more fine-grained measures. Therefore individual fairness notions are recommended in presence of both criteria (nodes 9 and 18).

Nodes 2, 3, and 11: In case one or more sources of bias are suspected ahead of time (before assessing fairness), the information can help warn against the use of some fairness notions. If representation bias is likely, the performance (accuracy) of the MLDM on under-represented categories will often be worse. Such disparity in performance between groups may lead to unreliable fairness assessment in case a group fairness notion is used, in particular disparate mistreatment notions (grayed section of the diagram). In such case, individual fairness notions can assess fairness more reliably provided that measurement bias is not likely (node 2). A suspicion of historical or measurement bias means that the features (X) and/or the label (Y) are not reliable. All group fairness notions using the label Y (disparate mistreatment) as well as individual notions are not recommended in that case. Statistical parity is recommended in such situation. Finally, in presence of either aggregation, evaluation, or deployment bias, causality-based fairness notions are recommended. The reason is that the interventional and counterfactual quantities used in the definitions of these notions go beyond mere correlations and hence allow to assess fairness more reliably in presence of such bias. For instance, Coston et al. [23] propose counterfactual formulations of fairness metrics to properly account for the effect of intervention (decision) on the outcome. Such effect is a type of deployment bias.

Node 3: As discussed in Section 5.12, there are several notions that use causal reasoning to assess fairness. Counterfactual fairness is suitable in case a fine-grained assessment is required as the equality of Eq. 21 conditions on all features (X). Counterfactual fairness, however, requires strong assumptions to be applicable in real scenarios (the availability of the full causal model including the latent variables distributions). Total effect (TE), effect of treatment on treated (ETT), and no proxy discrimination (nodes 13, 14 and 10), on the other hand, require a weaker assumption to be applicable, namely, the identifiability of the causal quantities used in their definitions. No proxy discrimination is recommended in presence of potential proxies, however, the identification of proxy variables requires a domain expertise of the application at hand. Finally, in case there are variables in the causal graph which are correlated with the sensitive attribute but in a manner that

is accepted as nondiscriminatory, no unresolved discrimination is recommended while the remaining causal based fairness notions should be avoided. No unresolved discrimination is easier to apply in practice as it only needs the availability of the causal graph.

Node 4: To reduce inequality and historical discrimination against sub-populations, in particular, minorities, some states and organizations resort to equality standards and regulations such as the laws enforced by the US Equal Employment Opportunity Commission [106]. In presence of such standards, to be deemed fair, an MLDM should satisfy such standards. Consequently, all what matters for fairness assessment is the proportion of positive prediction across all groups which corresponds to statistical parity.

Node 17: If no standards/regulations exist (node 4) and either the ground truth exists or the outcome label Y is available (node 11), statistical parity is not recommended (node 17) as it can lead to misleading results such as detecting unfairness in an otherwise fair scenario or failing to identify fairness in an unfair scenario. For instance, in stop-and-frisk real world scenario applied in New York city starting 1990 [80][27], the ground truth is available as by frisking an individual, a police officer can know with certainty the presence or no of illegal substance. In such case, one or several disparate mistreatment notions (nodes 30-41) are more suitable to assess fairness.

Nodes 22-47: The bulk of Figure 6 is dedicated for disparate mistreatment fairness notions and the criteria leading to each one of them. These notions define fairness in terms of the disparity of misclassification rates among the different groups in the population. Based on their definitions, selecting the most suitable notion to use depends on four citeria, namely, whether the emphasis is on precision or recall (node 22), whether the threshold is fixed or floating (nodes 23 and 24), whether the emphasis is on false negatives or false positives (nodes 26 and 28), and finally whether the emphasis is on the positive or negative class (node 27). As some notions focus only on either FP or FN (nodes 31, 32, 38, and 39), any notion that is insensitive to either FP or FN must not be used (nodes 42 - 47).

The diagram may be misleading if it is interpreted very categorically. This occurs when a user of the diagram navigates it and ends up using the recommended fairness notion without considering other important elements specific to the scenario at hand. The diagram can be misleading also when it is not clear which branch to take in a decision node. For example, the question in decision node

[27] Assuming the absence of measurement bias.

22 (emphasis on precision or recall?) is difficult to answer categorically in several scenarios. The decision nodes 4, 21, 12, and even 2, are typically easier to navigate, but can be challenging to settle in a number of scenarios. Moreover, in presence of measurement bias, the values of some features and even the outcome label may not be reliable which can make the diagram navigation more challenging. A potential solution would be to label one of the branches as default (to be followed when the answer is not clear), but this can, often result in a suboptimal decision. In summary, the diagram should be considered as guide and should never be used to supersede important elements specific to the scenario at hand.

Table 17: Correspondence between Fairness notions and the selection criteria: **C1**: disparate impact , **C2**: disparate treatment , **C3**: intersectionality/masking, **C4**: historical bias, **C5**: representational bias, **C6**: measurement bias, **C7**: aggregation/evaluation/deployment bias, **C8**: standards, **C9**: ground truth available, **C10**: y not reliable, **C11**: explanatory variables, **C12**: precision, **C13**: recall, **C14**: FP, **C15**: FN, **C16**: causal graph available, **C17**: threshold floating. Notation: ✓: recommended, ⚠: warning, ✗: must not, −: insensitive.

	Legal Frame			Suspected source of bias								Emphasis on		Emphasis on			
Criterion / Fairness notion	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17
Statistical parity	✓	⚠	⚠	✓	⚠	⚠	⚠	✓	⚠	✓	⚠	−	−	−	−	−	⚠
Conditional statistical parity	✓	⚠	⚠	✓	⚠	⚠	⚠	−	⚠	✓	✓	−	−	−	−	−	⚠
Equalized odds	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✓	✓	−	⚠
Equal opportunity	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✗	✓	−	⚠
Predictive equality	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✓	✓	−	⚠
Balance for positive class	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✗	✓	−	✓
Balance for negative class	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✓	✗	−	✓
Overall balance	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	⚠	✓	✓	✓	−	✓
Conditional use acc. equality	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	⚠	✓	✓	−	⚠
Predictive parity	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	⚠	✓	✗	−	⚠
Negative predictive parity	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	⚠	✗	✓	−	⚠
Calibration	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	⚠	−	−	−	✓
Well-calibration	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	⚠	−	−	−	✓
Overall accuracy equality	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	✓	✓	✓	✓	−	⚠
Treatment equality	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	−	−	✓	✓	−	−
Total fairness	✓	⚠	⚠	⚠	⚠	⚠	⚠	−	✓	⚠	⚠	−	−	✓	✓	−	⚠
Causal discrimination	✓	⚠	✓	⚠	✓	⚠	−	−	✓	⚠	⚠	−	−	−	−	−	−
Fairness through awareness	✓	⚠	✓	⚠	✓	⚠	⚠	−	✓	⚠	−	−	−	−	−	−	−
Total effect	−	✓	⚠	−	−	−	✓	−	−	−	−	−	−	−	−	✓	−
Effect of treatment on treated	−	✓	⚠	−	−	−	✓	−	−	−	−	−	−	−	−	✓	−
Counterfactual fairness	−	✓	✓	−	✓	−	✓	−	−	−	−	−	−	−	−	✓	−
No unresolved discrimination	−	✓	⚠	−	−	−	✓	−	−	−	✓	−	−	−	−	✓	−

Finally, Table 17 states explicitly the relationship between every selection criterion and every fairness notion. The table uses four symbols, namely, recommended (✓), warning (⚠), must-not (✗), and insensitive (−). Insensitive means that the choice of the fairness notion is independent of the selection criterion.

9. Conclusion

With the increasingly large number of fairness notions considered in the relatively new field of fairness in ML, selecting a suitable notion for a given MLDM (machine learning decision making) becomes a non-trivial task. There are two contributing factors. First, the boundaries between the defined notions are increasingly fuzzy. Second, applying inappropriately a fairness notion may report discrimination in an otherwise fair scenario, or vice versa, fail to identify discrimination in an unfair scenario. This survey tries to address this problem by identifying fairness-related characteristics of the scenario at hand and then use them to recommend and/or discourage the use of specific fairness notions. The main contribution of this survey is to systemize the selection process based on a decision diagram. Navigating the diagram will result in recommending and/or discouraging the use of fairness notions.

One of the main objectives of this survey is to bridge the gap between the real-world use case scenarios of automated (and generally unintentional) discrimination and the mostly technical tackling of the problem in the literature. Hence, the survey can be of particular interest to civil right activists, civil right associations, anti-discrimination law enforcement agencies, and practitioners in fields where automated decision making systems are increasingly used.

More generally, in real-scenarios, there are still two important obstacles to address the unfairness problem in automated decision systems. First, the victims of such systems are, very often, members of minority groups with limited influence in the public sphere. Second, automated decision systems are geared towards efficiency (typically money) and to optimize profit, they are designed to sacrifice the outliers as tolerable collateral damage. After all, the system is benefiting most of the population (employers finding ideal candidates, banks giving loans to minimum risk borrowers, a society with recidivists locked in prisons, etc.).

Acknowledgement

The work of Catuscia Palamidessi was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme. Grant agreement 835294.

[1] S. Lowry, G. Macpherson, A blot on the profession, British medical journal (Clinical research ed.) 296 (6623).

[2] C. O'Neill, Weapons of math destruction: How Big Data Increases Inequality and Threatens Democracy, Crown Publishers, 2016.

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

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

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

[6] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, in: C. H. Papadimitriou (Ed.), 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Vol. 67 of Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2017, pp. 43:1–43:23. doi:10.4230/LIPIcs.ITCS.2017.43.
URL http://drops.dagstuhl.de/opus/volltexte/2017/8156

[7] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 797–806.

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

[9] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big data 5 (2) (2017) 153–163.

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

[11] I. Zliobaite, A survey on measuring indirect discrimination in machine learning, arXiv preprint arXiv:1511.00148.

[12] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in criminal justice risk assessments: The state of the art, Sociological Methods & Research.

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

[14] A. Asuncion, D. Newman, Uci machine learning repository (2007).

[15] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, arXiv preprint arXiv:1908.09635.

[16] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 329–338.

[17] K. Makhlouf, S. Zhioua, C. Palamidessi, On the applicability of machine learning fairness notions, in: Bias and Fairness in AI Workshop at ECMLPKDD 2020, 2020.

[18] T. Kamishima, S. Akaho, J. Sakuma, Fairness-aware learning through regularization approach, in: 2011 IEEE 11th International Conference on Data Mining Workshops, IEEE, 2011, pp. 643–650.

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

[20] L. E. Celis, D. Straszak, N. K. Vishnoi, Ranking with fairness constraints, in: 45th International Colloquium on Automata, Languages, and Programming (ICALP 2018), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[21] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, A. Roth, Fairness in reinforcement learning, in: International Conference on Machine Learning, PMLR, 2017, pp. 1617–1626.

[22] J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, Prediction policy problems, American Economic Review 105 (5) (2015) 491–95.

[23] A. Coston, A. Mishler, E. H. Kennedy, A. Chouldechova, Counterfactual risk assessments, evaluation, and fairness, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 582–593.

[24] L. Weber, E. Dwoskin, Are workplace personality tests fair?, The Wall Street Journalhttps://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257.

[25] P. Lahoti, K. P. Gummadi, G. Weikum, ifair: Learning individually fair data representations for algorithmic decision making, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019, pp. 1334–1345.

[26] J. Leber, The machine-readable workforce, MIT Technology Reviewhttps://www.technologyreview.com/2013/05/27/178320/the-machine-readable-workforce.

[27] A. Waters, R. Miikkulainen, Grade: Machine learning support for graduate admissions, AI Magazine 35 (1) (2014) 64–64.

[28] M. V. Santelices, M. Wilson, Unfair treatment? the case of freedle, the sat, and the standardization approach to differential item functioning, Harvard Educational Review 80 (1) (2010) 106–134.

[29] K. Hao. The uk exam debacle reminds us that algorithms can't fix broken systems [online] (August 2020).

[30] COMPAS, Compas, https://www.equivant.com/northpointe-risk-need-assessments/ (2020).

[31] A. Majdara, M. R. Nematollahi, Development and application of a risk assessment tool, Reliability Engineering & System Safety 93 (8) (2008) 1130–1137.

[32] J. R. Meyers, F. Schmidt, Predictive validity of the structured assessment for violence risk in youth (savry) with juvenile offenders, Criminal Justice and Behavior 35 (3) (2008) 344–355.

[33] predPol, predpol, https://www.predpol.com (2020).

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

[35] A. Chalfin, O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, S. Mullainathan, Productivity and selection of human capital with machine learning, American Economic Review 106 (5) (2016) 124–27.

[36] M. Rhee, Impact: The dcps evaluation and feedback system for school-based personnel, https://dcps.dc.gov/page/impact-dcps-evaluation-and-feedback-system-school-based-personnel (2019).

[37] K. Quick, The unfair effects of impact on teachers with the toughest jobs, The Century Foundationhttps://tcf.org/content/commentary/the-unfair-effects-of-impact-on-teachers-with-the-toughest-jobs/?agreed=1.

[38] R. Vaithianathan, T. Maloney, E. Putnam-Hornstein, N. Jiang, Children in the public benefit system at risk of maltreatment: Identification via predictive modeling, American journal of preventive medicine 45 (3) (2013) 354–359.

[39] V. Eubanks, Automating inequality: How high-tech tools profile, police, and punish the poor, St. Martin's Press, 2018.

[40] E. K. Spanakis, S. H. Golden, Race/ethnic difference in diabetes and diabetic complications, Current diabetes reports 13 (6) (2013) 814–823.

[41] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (6464) (2019) 447–453.

[42] H. Suresh, J. V. Guttag, A framework for understanding unintended consequences of machine learning, arXiv preprint arXiv:1901.10002.

[43] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, Recommender systems: an introduction, Cambridge University Press, 2010.

[44] A. Lambrecht, C. E. Tucker, Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads, An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads (March 9, 2018).

[45] A. Datta, M. C. Tschantz, A. Datta, Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination, Proceedings on privacy enhancing technologies 2015 (1) (2015) 92–112.

[46] Google, Google ad setting, `http://www.goolge.com/settings/ads`.

[47] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.

[48] A. Dehghan, E. G. Ortiz, G. Shu, S. Z. Masood, Dager: Deep age, gender and emotion recognition using convolutional neural network, arXiv preprint arXiv:1702.04280.

[49] C. Fabian Benitez-Quiroz, R. Srinivasan, A. M. Martinez, Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5562–5570.

[50] R. Srinivasan, J. D. Golomb, A. M. Martinez, A neural basis of facial action recognition in humans, Journal of Neuroscience 36 (16) (2016) 4434–4442.

[51] X. Wu, X. Zhang, Automated inference on criminality using face images, arXiv preprint arXiv:1611.04135 (2016) 4038–4052.

[52] C. Garvie, The perpetual line-up: Unregulated police face recognition in America, Georgetown Law, Center on Privacy & Technology, 2016.

[53] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on fairness, accountability and transparency, 2018, pp. 77–91.

[54] Y. R. Shrestha, Y. Yang, Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems, Algorithms 12 (9) (2019) 199.

[55] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, X. Zhu, Employee turnover prediction with machine learning: A reliable approach, in: Proceedings of SAI intelligent systems conference, Springer, 2018, pp. 737–758.

[56] A. M. Esmaieeli Sikaroudi, R. Ghousi, A. Sikaroudi, A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing), Journal of Industrial and Systems Engineering 8 (4) (2015) 106–121.

[57] R. S. Sexton, S. McMurtrey, J. O. Michalopoulos, A. M. Smith, Employee turnover: a neural network solution, Computers & Operations Research 32 (10) (2005) 2635–2651.

[58] D. Alao, A. Adeyemo, Analyzing employee attrition using decision tree algorithms, Computing, Information Systems, Development Informatics and Allied Research Journal 4.

[59] P. T. M. Marope, P. J. Wells, E. Hazelkorn, Rankings and accountability in higher education: Uses and misuses, Unesco, 2013.

[60] R. A. QC, D. Masters, Regulating For an Equal AI: A New Role for Equality Bodies, EQUINET: the European Network of Equality Bodies, 2020, iSBN: 978-92-95112-35-3.
URL https://equineteurope.org/wp-content/uploads/2020/06/ai_report_digital.pdf

[61] M. K. Dodson, W. A. Cliby, G. L. Keeney, M. F. Peterson, K. C. Podritz, Skene's gland adenocarcinoma with increased serum level of prostate-specific antigen, Gynecologic oncology 55 (2) (1994) 304–307.

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

[63] K. Crenshaw, Mapping the margins: Intersectionality, identity politics, and violence against women of color, Stan. L. Rev. 43 (1990) 1241.

[64] S. Barocas, A. D. Selbst, Big data's disparate impact, Calif. L. Rev. 104 (2016) 671.

[65] V. Zarya, The share of female ceos in the fortune 500 dropped by 25% in 2018, https://fortune.com/2018/05/21/women-fortune-500-2018/ (2018).

[66] K. Crawford, Think again: big data. why the rise of machines isn't all it's cracked up to be, Foreign Policy 10.

[67] Y.-M. Li, C. Peng, J.-G. Zhang, W. Zhu, C. Xu, Y. Lin, X.-Y. Fu, Q. Tian, L. Zhang, Y. Xiang, et al., Genetic risk factors identified in populations of european descent do not improve the prediction of osteoporotic fracture and bone mineral density in chinese populations, Scientific reports 9 (1) (2019) 1–9.

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

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

[70] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd innovations in theoretical computer science conference, 2012, pp. 214–226.

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

[72] S. Barocas, M. Hardt, A. Narayanan, Fairness in machine learning, NIPS Tutorial.

[73] I. Zliobaite, On the relation between accuracy and fairness in binary classification, arXiv preprint arXiv:1505.05723.

[74] C. Simoiu, S. Corbett-Davies, S. Goel, et al., The problem of infra-marginality in outcome tests for discrimination, The Annals of Applied Statistics 11 (3) (2017) 1193–1216.

[75] T. J. VanderWeele, M. A. Hernán, Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs, American journal of epidemiology 175 (12) (2012) 1303–1310.

[76] J. E. Johndrow, K. Lum, et al., An algorithm for removing sensitive information: application to race-independent recidivism prediction, The Annals of Applied Statistics 13 (1) (2019) 189–220.

[77] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Advances in neural information processing systems, 2016, pp. 3315–3323.

[78] F. Kamiran, I. Zliobaite, T. Calders, Quantifying explainable discrimination and removing illegal discrimination in automated decision making, Knowledge and information systems (Print) 35 (3) (2013) 613–644.

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

[80] J. Bellin, The inverse relationship between the constitutionality and effectiveness of new york city stop and frisk, BUL Rev. 94 (2014) 1495.

[81] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions, The quarterly journal of economics 133 (1) (2018) 237–293.

[82] P. Garg, J. Villasenor, V. Foggo, Fairness metrics: A comparative analysis, arXiv preprint arXiv:2001.07864.

[83] S. Galhotra, Y. Brun, A. Meliou, Fairness testing: testing software for discrimination, in: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, 2017, pp. 498–510.

[84] M. S. Nikulin, Hellinger distance, Encyclopedia of mathematics 78.

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

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

[87] K. A. Bollen, Structural equations with latent variables wiley, New York.

[88] J. Pearl, Causality, Cambridge university press, 2009.

[89] I. Shpitser, J. Pearl, Complete identification methods for the causal hierarchy, Journal of Machine Learning Research 9 (Sep) (2008) 1941–1979.

[90] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, Avoiding discrimination through causal reasoning, in: Advances in Neural Information Processing Systems, 2017, pp. 656–666.

[91] J. Pearl, Direct and indirect effects, in: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence, 2001, pp. 411–420.

[92] A. Khademi, S. Lee, D. Foley, V. Honavar, Fairness in algorithmic decision making: An excursion through the lens of causality, in: The World Wide Web Conference, 2019, pp. 2907–2914.

[93] J. Zhang, E. Bareinboim, Fairness in decision-making—the causal explanation formula, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[94] J. Zhang, E. Bareinboim, Equality of opportunity in classification: A causal approach, in: Advances in Neural Information Processing Systems, 2018, pp. 3671–3681.

[95] S. Chiappa, Path-specific counterfactual fairness, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 7801–7808.

[96] Y. Wu, L. Zhang, X. Wu, H. Tong, Pc-fairness: A unified framework for measuring causality-based fairness, in: Advances in Neural Information Processing Systems, 2019, pp. 3404–3414.

[97] E. H. Simpson, The interpretation of interaction in contingency tables, Journal of the Royal Statistical Society: Series B (Methodological) 13 (2) (1951) 238–241.

[98] P. J. Bickel, E. A. Hammel, J. W. O'Connell, Sex bias in graduate admissions: Data from berkeley, Science 187 (4175) (1975) 398–404.

[99] J. S. Kim, J. Chen, A. Talwalkar, Model-agnostic characterization of fairness trade-offs, arXiv preprint arXiv:2004.03424.

[100] M. B. Zafar, I. Valera, M. G. Rodriguez, K. P. Gummadi, Fairness constraints: Mechanisms for fair classification, arXiv preprint arXiv:1507.05259.

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

[102] G. Yona, G. Rothblum, Probably approximately metric-fair learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 5680–5688.

[103] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: International Conference on Machine Learning, PMLR, 2018, pp. 2564–2572.

[104] S. Wachter, B. Mittelstadt, A right to reasonable inferences: re-thinking data protection law in the age of big data and ai, Colum. Bus. L. Rev. (2019) 494.

[105] R. Binns, On the apparent conflict between individual and group fairness, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 514–524.

[106] Us equal employment opportunity commission, `https://www.eeoc.gov/`.

[107] P. Judea, Causality: models, reasoning, and inference, Cambridge University Press. ISBN 0 521 (77362) (2000) 8.

Appendix A. Counterfactual probability computation using the three-step procedure

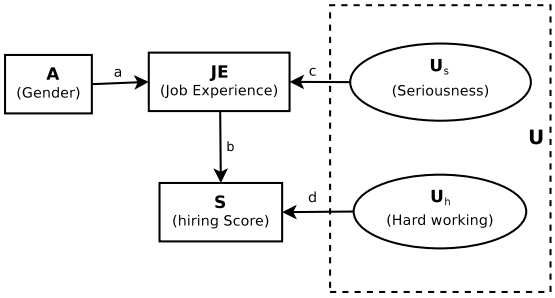

Figure A.7: A simple deterministic causal graph for the hiring example.

The probability of the counterfactual realization $P(\hat{Y}_{A\leftarrow a_1} \mid X = x, A = a_0)$ is computed using the following three-steps process [107]:

1. **Abduction**: update the probability $P(U = u)$ given the evidence to obtain: $P(U = u \mid X = x, A = a_0)$.
2. **Action**: set the sensitive attribute value A to a_1 and update all structural functions of the causal graph accordingly.

3. **Prediction**: compute the outcome ($\hat{Y}$) value using the updated probability $P(U \mid X = x, A = a_0)$ and structural functions.

To illustrate how counterfactual quantities are computed, consider the simplified deterministic version of the hiring example in Figure A.7. For simplicity, the hiring score variable S depends on the observable variable JE representing job experience and the exogenous variable U_h representing how hard working the candidate is. The variable JE in turn depends on the observable sensitive variable A representing the gender (male or female) and the exogenous variable U_s representing the seriousness of the candidate. The causal graph in Figure A.7 is represented by the two following equations:

$$JE = a.A + c.U_s \tag{A.1}$$

$$S = b.JE + d.U_h \tag{A.2}$$

For simplicity of the illustration, assume that both U (U_s and U_h) variables are independent and all the parameters of the model (Eq. A.1 and A.2) are known. Assume that the values of the coefficients are given as follows:

$$a = 0.1, \quad b = 0.7, \quad c = 0.9, \quad d = 0.3$$

Given this causal model, consider a candidate John who is male ($A^{John} = 1$), with the normalized[28] job education level $JE^{John} = 0.6$ and a predicted score $\hat{S}^{John} = 0.55$. Assessing the fairness of the hiring score prediction with respect to gender is achieved through answering the following question: *what would John's hiring score have been had he was of opposite gender (female)?* This corresponds to the hiring score of John in the counterfactual world where John is a female ($\hat{S}^{John}_{A \leftarrow 0}$). To compute this quantity, the three-steps process above is used, namely, abduction, action, and prediction.

The abduction step consists in using the evidence ($A^{John} = 1, JE^{John} = 0.6, \hat{S}^{John} = 0.55$) to identify the specific characteristics of $John$, namely, his level of seriousness and hard working (U_s

[28] To keep the computation simple, all variable values are normalized between 0 and 1.

and U_h)[29] as follows:

$$\begin{aligned} U_s^{John} &= \frac{JE^{John} - a.A^{John}}{c} \\ &= \frac{5}{9} \\ U_h^{John} &= \frac{\hat{S}^{John} - b.JE^{John}}{d} \\ &= \frac{13}{30} \end{aligned} \tag{A.3}$$

The second step consists in setting the sensitive attribute A^{John} to the opposite gender (0) and updating all equations of the model. This consists in replacing the variable A in Eq. A.1 by 0.

The third step consists in the prediction, that is computing $\hat{S}_{A\leftarrow 0}$ in the counterfactual world. This requires the computation of $JE_{A\leftarrow 0}^{John}$, that is, the job experience of John in a world where John is a female.

$$\begin{aligned} JE_{A\leftarrow 0}^{John} &= a.0 + c.U_s^{John} \\ &= 0.5 \end{aligned} \tag{A.4}$$

$$\begin{aligned} \hat{S}_{A\leftarrow 0}^{John} &= b.JE_{A\leftarrow 0}^{John} + d.U_h^{John} \\ &= 0.48 \end{aligned} \tag{A.5}$$

Hence, the hiring score of John had he was female is $\hat{S}_{A\leftarrow 0}^{John} = 0.48$ which is considered a violation of counterfactual fairness as the predicted hiring score of John in the original world is $\hat{S}^{John} = 0.55$.

Consider now a female candidate Marie ($A^{Marie} = 0$), with the a job education level $JE^{Marie} = 0.61$ and a predicted score $\hat{S}^{Marie} = 0.65$. The question to investigate is now: *what would Marie's hiring score have been had she was male?* This boils down to computing $\hat{S}_{A\leftarrow 1}^{Marie}$ and comparing it with $\hat{S}^{Marie} = 0.65$. Applying the three-steps process:

[29] Since this example is deterministic, every individual is characterized by a unique assignment for exogenous variables U_s and U_h. In typical (non-deterministic) scenarios, every individual is assigned a probability distribution over the exogenous variables.

Abduction:

$$\begin{aligned} U_s^{Marie} &= \frac{JE^{Marie} - a.A^{Marie}}{c} \\ &= \frac{61}{90} \\ U_h^{Marie} &= \frac{\hat{S}^{Marie} - b.JE^{Marie}}{d} \\ &= \frac{223}{30} \end{aligned} \tag{A.6}$$

Action: replacing the variable A in Eq. A.1 by 1.

Prediction:

$$\begin{aligned} JE_{A\leftarrow 1}^{Marie} &= a.1 + c.U_s^{Marie} \\ &= 0.71 \end{aligned} \tag{A.7}$$

$$\begin{aligned} \hat{S}_{A\leftarrow 1}^{Marie} &= b.JE_{A\leftarrow 1}^{Marie} + d.U_h^{Marie} \\ &= 0.72 \end{aligned} \tag{A.8}$$

$\hat{S}_{A\leftarrow 1}^{Marie} = 0.72 > \hat{S}^{Marie} = 0.65$ is another violation for counterfactual fairness.