# On the Applicability of Machine Learning Fairness Notions

Karima Makhlouf
Université du Québec à
Montréal
Montréal, Canada
makhlouf.karima@courrier.uqam.ca

Sami Zhioua
Higher Colleges of Technology
Dubai, UAE
szhioua@hct.ac.ae

Catuscia Palamidessi
INRIA, École Polytechnique,
IPP
Paris, France
catuscia@lix.polytechnique.fr

## ABSTRACT

Machine Learning (ML) based predictive systems are increasingly used to support decisions with a critical impact on individuals' lives such as college admission, job hiring, child custody, criminal risk assessment, etc. As a result, fairness emerged as an important requirement to guarantee that 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 of fairness notions that, unlike other surveys in the literature, 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.

## 1. INTRODUCTION

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 its 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 [32; 36]. 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.

In the context of automated decision-making, a consensual definition of fairness can be formulated as: "*absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits*" [34]. Mathematically, however, there is no consensual definition of fairness. Very often, research papers focus on a specific real-world scenario of automated decision system and propose a fairness definition tailored to that scenario and its specificities. Consequently, several fairness notions have been introduced in the literature. These notions are the subject of several survey papers [2; 6; 19; 34; 35; 41; 44].

The very reason of having different flavors of fairness notions is how suitable each one of them is for specific real-world scenarios. But none of the existing surveys addressed this aspect. Discussion about the suitability (and sometimes the applicability) of the fairness notions is very limited and scattered through several papers [3; 10; 29; 35; 43]. In this survey paper we show that each ML-based automated decision 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.

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 ML-based automated decision 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.

## 2. 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 unacceptably 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

[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.

than others.

***Job hiring***: MLDMs in hiring are increasingly used by employers to automatically screen candidates for job openings.Typically, the input data used by the MLDM include: affiliation, education level, job experience, IQ score, age, gender, address, etc. The MLDM outputs a decision and/or a score indicating how 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.). This 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[2]. 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 [18]. Unfair college admission MLDMs may discriminate against a certain ethnic group (e.g. African-American [38]) which could lead, in the long term, to economic inequalities and corrupting the role of higher education in society as a whole.

***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). 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 [1], 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.

***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. 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 get 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 [17].

***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 disproportionally higher rate than the dominant population. For instance, diabetic patients have known differences in associated complications across ethnicities [40].

***Facial analysis***: Automated facial analysis systems are used to identify perpetrators from security video footages, to detect melanoma (skin cancer) from face images [16], to detect emotions [12], and to even determine individual's characteristics such as IQ, propensity towards terrorist crime, etc. based on their face images [42]. 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 footages) 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 [22].

***Others***: Other MLDMs with fairness concerns include: insurance policy prediction [39], income prediction [34], teachers evaluation and promotion [7], online recommendation [25] and university ranking [33; 36].

# 3. 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. 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. We tried to be exhaustive when listing the decision criteria based on the existing literature. However, there are no guarantees about the completeness of this list.

***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

[2] 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.

the individual to a blood test[3] 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.

***Base rate is the same across groups***: The base rate is the proportion of positive outcome in a population (Based on Table 2, $BR = \frac{TP+FN}{TP+FP+TN+FN}$). This 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[4].

***(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 explaining variables***: An explaining variable[5] 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 rate 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 [13]) is defined as the fraction of positive instances among the predicted positive instances ($\frac{TP}{TP+FP}$). In other words, if the system predicts an instance as positive, how precise that prediction is. Recall (the complement of model error [13]) is defined as the fraction of the total number of positive instances that are correctly predicted positive ($\frac{TP}{TP+FN}$). In other words, how many of the positive instances the system is able to identify. There is always a tradeoff 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 contribute more to fairness than making sure that an applicant who has been granted a loan really deserves it. 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 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[6]. 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.

***Likelihood of intersectionality***: Intersectionality theory [11] 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 [4]. 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".

***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 [3]. Another example might be an internal organizational policy imposing diversity among its employees.

## 4. 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 $Y$ and $\hat{Y}$ be binary random variables representing, respectively, the actual outcome and the predicted outcome where $Y = 1$ designates a positive instance, while $Y = 0$ a negative one. Typically, the predicted outcome $\hat{Y}$ is derived from a score represented by a random variable $S$ where $P(S = s)$ is

[3] assuming the blood test is flawless.

[4] While male prostate cancer is the second most common cancer in men, female prostate cancer is rare [14].

[5] Referred also as resolving variable.

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

the probability that the score value is equal to $s$.

All fairness notions presented in this survey (Table 1) 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 is it.

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*[7]. 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.

***Statistical parity***[15]: is one of the most commonly accepted notions of fairness. It requires the prediction to be statistically independent of the sensitive attribute ($\hat{Y} \perp A$). In other words, the predicted acceptance rates for both protected and unprotected groups should be equal. Using the confusion matrix (Table 2), statistical parity implies that $\frac{TP+FP}{TP+FP+FN+TN}$ is equal for both groups. Statistical parity is appealing in scenarios where there is a preferred decision over the other. For example, being accepted to a job, not being arrested, being admitted to a college, etc.[8]. 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 the ideal case ($\hat{y} = y$), this will lead to loss of utility [24]. 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* [15] 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 [4]. Masking is possible to game several fairness notions, but it is particularly easy to carry out in the case of statistical parity.

***Conditional statistical parity*** [10]: this notion is a variant of statistical parity obtained by controlling on a set of legitimate attributes[9]. 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* discrimination. In other words, this notion removes the illegal discrimination, allowing the disparity in decisions to be present as long as they are explainable [10]. 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. 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 [26].

***Equalized odds*** [23]: this notion 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 that both sub-populations to have the same $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$ (Table 2). 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 [5]. 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. Because equalized odds requirement is rarely satisfied in practice, two variants can be obtained by relaxing its equation (Table 1). The first one is called **equal opportunity** [23] and is obtained by requiring only TPR equality among groups. As $TPR$ 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 second relaxed variant of equalized odds is called **predictive equality** [10] which requires only the FPR to be equal in both groups. Since $FPR$ is independent from $FN$, predictive equality is completely insensitive to false negatives. Predictive equality is particularly suitable to measure the fairness of face recognition systems in crime investigation where security camera footages 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 ethnic 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.

***Conditional use accuracy equality*** [6]: with this notion,

[7] Known also as: blindness, unawareness [35], anti-classification [9], and treatment parity [31].

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

[9] Called explanatory attributes in [26].

Table 1: Classification of fairness notions.

| Fairness Notion | Reference | Formulation | Classification | Type |
|---|---|---|---|---|
| Statistical Parity | [15] | $P(\hat{Y} \mid A=0)=P(\hat{Y} \mid A=1)$ | Independence<br>$\hat{Y} \perp A$ | Group Fairness |
| Conditional Statistical Parity | [10] | $P(\hat{Y}=1 \mid E=e, A=0)=P(\hat{Y}=1 \mid E=e, A=1) \quad \forall e$ | | |
| Equalized Odds | [23] | $P(\hat{Y}=1 \mid Y=y,\ A=0)=P(\hat{Y}=1 \mid Y=y,\ A=1) \quad \forall y \in \{0,1\}$ | Separation<br>$\hat{Y} \perp A \mid Y$ | |
| Equal Opportunity | | $P(\hat{Y}=1 \mid Y=1, A=0)=P(\hat{Y}=1 \mid Y=1, A=1)$ | | |
| Predictive Equality | [10] | $P(\hat{Y}=1 \mid Y=0, A=0)=P(\hat{Y}=1 \mid Y=0, A=1)$ | | |
| Balance for Positive Class | [29] | $E[S \mid Y=1, A=0)]=E[S \mid Y=1, A=1]$ | | |
| Balance for Negative Class | | $E[S \mid Y=0, A=0]=E[S \mid Y=0, A=1]$ | | |
| Conditional Use Accuracy Equality | [6] | $P(Y=y \mid \hat{Y}=y, A=0)=P(Y=y \mid \hat{Y}=y, A=1) \quad \forall y \in \{0,1\}$ | Sufficiency<br>$Y \perp A \mid \hat{Y}$ | |
| Predictive Parity | [8] | $P(Y=1 \mid \hat{Y}=1, A=0)=P(Y=1 \mid \hat{Y}=1, A=1)$ | | |
| Calibration | | $P(Y=1 \mid S=s, A=0)=P(Y=1 \mid S=s, A=1) \quad \forall s \in [0,1]$ | | |
| Well-calibration | [29] | $P(Y=1 \mid S=s, A=0)=P(Y=1 \mid S=s, A=1)=s \quad \forall\, s \in [0,1]$ | | |
| Overall Accuracy Equality | [6] | $P(\hat{Y}=Y \mid A=0)=P(\hat{Y}=Y \mid A=1)$ | Other metrics<br>from confusion matrix | |
| Treatment Equality | | $\frac{FN}{FP}{}_{(\mathrm{a=0})}=\frac{FN}{FP}{}_{(\mathrm{a=1})}$ | | |
| Total Fairness | | - | Independence,<br>Separation and Sufficiency | |
| No unresolved discrimination | [27] | - | Causality | |
| 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 | [30] | $P(\hat{Y}_{A\leftarrow a}(U)=y \mid X=x, A=a)=P(\hat{Y}_{A\leftarrow a'}(U)=y \mid X=x, A=a)$ | | Individual Fairness |
| Causal Discrimination | [20] | $X_{(\mathrm{a=0})}=X_{(\mathrm{a=1})} \ \wedge\ A_{(\mathrm{a=0})} \neq A_{(\mathrm{a=1})} \ \Rightarrow \hat{y}_{(\mathrm{a=0})}=\hat{y}_{(\mathrm{a=1})}$ | Similarity Metric | |
| Fairness Through Awareness | [15] | $D(M(v_i), M(v_j)) \leq d(v_i, v_j)$ | | |

Table 2: Confusion matrix

| | Actual Positive<br>$Y = 1$ | Actual Negative<br>$Y = 0$ |
|---|---|---|
| Predicted<br>Positive<br>$\hat{Y} = 1$ | **TP**<br>(True Positive) | **FP**<br>(False Positive)<br>*Type I error* |
| Predicted<br>Negative<br>$\hat{Y} = 0$ | **FN**<br>(False Negative)<br>*Type II error* | **TN**<br>(True Negative) |

fairness 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. By contrast to equalized odds, one is conditioning on the algorithm's predicted outcome not the actual outcome. In other words, the emphasis is on the precision of the MLDM system rather than its recall. **Predictive parity** [8] is a relaxation of conditional use accuracy equality requiring only equal $PPV$ among groups. Like predictive equality, predictive parity is insensitive to false negatives. Hence in any scenario where fairness is sensitive to false negatives, predictive parity should not be used. 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 [43]. 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.

***Balance*** [29] : 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. **Balance for positive class** focuses on the individuals who constitute positive instances and is satisfied if the average score $S$ received by those individuals is the same for both groups. The intuition behind this notion is that a balance for the positive class should be assured, thus, a violation of this balance means that individuals belonging to the positive class in one group might receive steadily lower predicted score than individuals belonging to the positive class in the other group. **Balance of negative class** is an analogous fairness notion where the focus is on the negative class. Both variants of balance can be required simultaneously which leads to a stronger notion of balance[10]. 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).

***Calibration*** [8]: 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. Interestingly, calibration is not always stronger than predictive parity [21]. 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 where the threshold for intervention (withdrawing a child from his family) depends on the available seats in foster houses. **Well-calibration** [29] 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.

***No unresolved discrimination*** [27]: similarly to counterfactual fairness, no unresolved discrimination is assessed using causal reasoning. Given a causal graph, no unresolved discrimination is satisfied when no directed path 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 it is accepted as nondiscriminatory (this is very similar to the use of the explanatory attributes in conditional statistical parity but in a non-causal context). 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. The application of unresolved discrimination is completely based on the definition of the causal graph. Thus, this notion is well-suited when a reliable and trustworthy causal graph that describes best the domain at hand including all relevant relations and features (in particular, the resolving attributes) is available. Hence, it is mandatory that the choice of the resolving variables along with their causal relationships to the other attributes is in reliance on policy makers and domain professionals expertise.

***No proxy discrimination*** [27]: a causal graph exhibits potential proxy discrimination 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 merely 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 the equality of the following equation is valid for all potential proxies $P_x$:

$$P(\hat{Y} \mid do(P_x = p)) = P(\hat{Y} \mid do(P_x = p')) \quad \forall\, p, p'$$

In other words, this notion implies that changing the value of $P_x$ should not have any impact on the prediction. As with the previous two fairness notions, the applicability of proxy discrimination is based on the construction of a reliable and plausible causal graph. In particular, the main goal of this notion is to carefully investigate and analyze the relations between attributes (in particular, those related to the sensitive attributes) in order to discover all potential proxies that might result in unfair decisions.

***Counterfactual fairness*** [30] : counterfactual is a con-

[10] No previous work reported such fairness notion.

cept from causal inference which goes beyond mere statistical correlation between variables and relies on causal relationship between them. Causal relationships are represented using causal graph where nodes represent variables (attributes) and edges represent causal relationships between variables. $U$ represents all *exogenous* variables such that each assignment $U = u$ corresponds to a unique individual in the population or to a situation in nature [37]. Counterfactual fairness is achieved if for every individual $(U = u, X = x, A = a)$ of the entire population, the probability to be predicted as hired is the same, *had $A$ been $a'$*. That is, $P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a)$ is equal to $P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)$ where $\hat{Y}_{A\leftarrow a'}(U)$ is called the counterfactual and corresponds to the predicted outcome in case the variable $A$ is *forced*[11] to be equal to $a'$ for an individual with exogenous variable $U = u$. The probability of the counterfactual is conditioned on $(X = x, A = a)$ which is called the evidence. In other words, counterfactual fairness requires that for every individual $(X = x, A = a)$ in the population (evidence), the probability of the outcome is the same in both the actual world $(A \leftarrow a)$ and the counterfactual world $(A \leftarrow a')$. Compared to causal discrimination where all variables are measured in the same world but on different individuals, counterfactual fairness measures variables on the same individual but in different worlds (the world of the evidence, and another hypothetical world). This notion is satisfied if the probability distribution of the predicted outcome $\hat{Y}$ is the same in the actual and counterfactual worlds, for every possible individual. A simple but important implication of counterfactual fairness formulation 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 $A$. Consequently, one can tell if a predictor is counterfactually fair by simply checking the causal graph[12]. Hence, the main challenge to using counterfactual fairness in practice is the construction of the causal graph which typically requires domain expertise. It is important to note that generally, data can be used to validate a proposed causal graph. That is, a dataset of observed samples can be used to rule out possible causal graphs.

***Causal discrimination*** [20]: this notion implies that a classifier should produce exactly the same prediction for individuals who differ only in their sensitive attribute $A$ while possessing identical attributes $X$. At a first glance, causal discrimination can be seen as an extreme case of conditional statistical parity 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 the mathematical formulation of conditional statistical parity (Table 1) is expressed in terms of conditional probabilities. Causal discrimination, however, considers every individual separately regardless of its contribution to sub-population proportions. 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. The result of applying causal discrimination is the percentage of violations in the entire population (i.e. how many individuals are unfairly treated).

***Fairness through awareness*** [15]: this notion 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 $j$:

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

In practice, fairness through awareness assumes that the similarity metric is known for each pair of individuals [28]. 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 [30].

## 5. DIAGRAM AND DISCUSSION

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 3) 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 1. 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 1 is designed such that every node indicates which notions are recommended, which notions should be avoided, and which notions must not be used.

The diagram is composed of four types of nodes:

- **Decision node (diamond):** based on fairness-related characteristics (Section 3)
- **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.

[11] Pearl et al. [37] use the term *surgical modification.*

[12] Kusner et al. [30] identify some exceptions, but guaranteeing that they will *not happen in general.*

To illustrate how the diagram should be interpreted, consider the recommended node predictive parity (34). According to the diagram, predictive parity is recommended in the scenario where intersectionality and/or masking are unlikely (decision node 1), standards do not exist (decision node 2), ground-truth is available or outcome $Y$ is reliable (decision node 6), fairness is more sensitive to precision rather than recall (decision node 14), the prediction threshold is typically fixed (decision node 20) and the emphasis is on false positives rather than false negatives (decision node 24). In that particular scenario, equal opportunity must not be used (must-not node 42) because fairness in this scenario is particularly sensitive to false positives, while equal opportunity is completely insensitive to false positives. The warning node 9 along the same path indicates that statistical parity is not suitable in this scenario. Finally, any fairness notion for which there is no warning node or must-not node along the path of the scenario can be used in this scenario. For instance, all individual fairness notions can be used, which is indicated by the link to node 4, i.e., the square with a "4" inside at the end of several paths, as will be discussed below. The lower part of the diagram corresponding to the "yes" branch of decision node 1 deals with individual fairness notions. In that branch all group fairness notions are not recommended (warning node 3) because they are not suitable when intersectionality or masking are likely.The part between decision nodes 7 and 15 is the only part with a non-tree-like structure. It expresses the fact that, typically, several individual fairness notions can be suitable at the same time. This indicates also that currently, the tensions between the various individual fairness notions are not well understood in the literature.

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 14 (emphasis on precision or recall?) is difficult to answer categorically in several scenarios. The decision nodes 13, 19, 22, and even 1, are typically easier to navigate, but can be challenging to settle in a number of scenarios. 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.

## 6. 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. Hence, the survey is an attempt 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.

## 7. AKNOWLEDGMENTS

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.

## 8. REFERENCES

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

[2] Barocas, S., Hardt, M., Narayanan, A.: Fairness in machine learning. NIPS Tutorial (2017)

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

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

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

[6] Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research p. 0049124118782533 (2018)

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

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

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

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

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

[12] Dehghan, A., Ortiz, E.G., Shu, G., Masood, S.Z.: Dager: Deep age, gender and emotion recognition using convolutional neural network. arXiv:1702.04280 (2017)

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

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

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

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

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

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

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

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

[21] Garg, P., Villasenor, J., Foggo, V.: Fairness metrics: A comparative analysis. arXiv preprint arXiv:07864 (2020)

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

[23] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. arXiv preprint arXiv:02413 (2016)

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

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

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

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

[28] Kim, M., Reingold, O., Rothblum, G.: Fairness through computationally-bounded awareness. In: NIPS. pp. 4842–4852 (2018)

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

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

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

[32] Lowry, S., Macpherson, G.: A blot on the profession. British medical journal (Clinical research ed.) **296**(6623), 657 (1988)

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

[34] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)

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

[36] O'Neill, C.: Weapons of math destruction. How Big Data Increases Inequality and Threatens Democracy (2016)

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

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

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

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

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

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

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

[44] Zliobaite, I.: A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148 (2015)

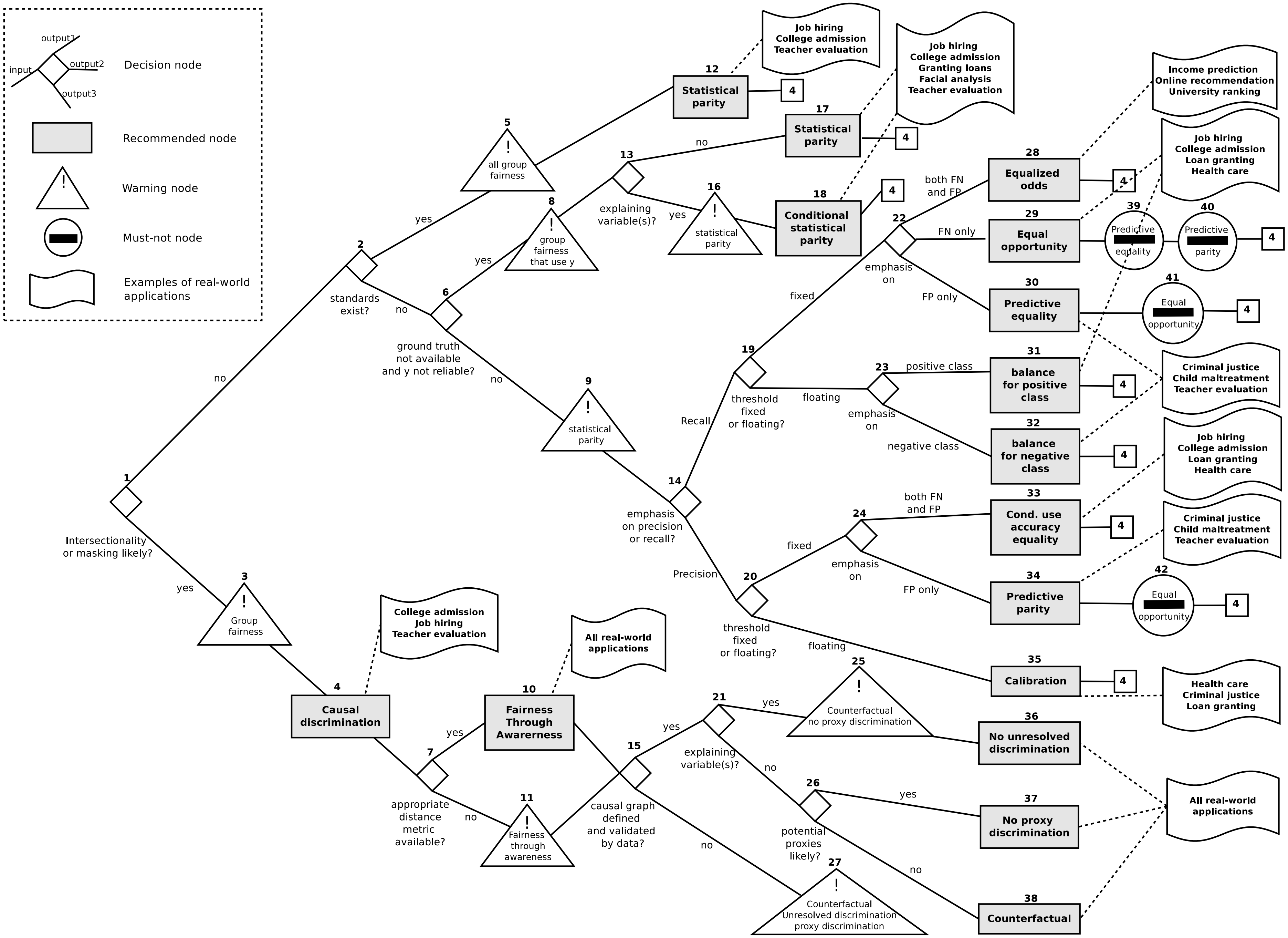

Figure 1: Fairness notions applicability decision diagram