





ARTICLE

Fairness definitions explained and illustrated with examples

Pavlo Burda  and Sieuwert van Otterloo 

Email: sieuwert@ictinstitute.nl

(First published online August 30, 2025)

Abstract

Algorithms are increasingly used to make or support decisions, such as recruitment and selection, loan approval, fraud detection, classification and prediction. This has raised important questions on how to assure the use of algorithms respects important values such as fairness, absence of bias and discrimination, and also explainability and redress. Despite decades of research, there is no single metric that can be used to measure fairness, absence of bias or discrimination. For every implementation one needs to evaluate multiple different metrics and also additional decisions about the subgroups and thresholds to use for the fairness metrics. In this paper we present an overview of the most important fairness metrics, provide, updated more thorough definitions, and a new dataset of hiring candidates ($n=225$) to show the impact of the different definition in practice. The dataset presented is suitable for educational and demonstration purposes since it is freely usable, has no privacy constraints, no data quality issues and is sufficiently complex to explain the challenges of fairness evaluations. We use the dataset to show that the different fairness definitions are indeed conflicting and that it is necessary to consider many subgroups using multiple sensitive features. We also discuss the importance of the socio-technical context and domain knowledge in fairness.

1. Introduction

Machine learning has become prevalent in many different domains. The use of machine learning enables making decisions based on relevant data. The benefit of automated decision-making is that it allows for more efficiency and accuracy when making a decision (with less subjectivity and less human biases when making a decision), as it is able to consider a larger pool of relevant factors that could affect a decision [CCG⁺22]. However, the use of machine learning for decision-making in practice, has shown to have the potential to discriminate unlawfully and undesirably, by replicating models of discrimination and bias found in both algorithms and data. Therefore the collection and selection of training data may be a predominant means of introducing bias. Without sufficiently inclusive and balanced training data, the system may learn to make unfair decisions, and therefore create algorithmic biases [CLM22].

One of the best known efforts to tackle the power of algorithmic systems trends, with fairness as one of the core principles, is the introduction of EU's GDPR in 2016 [Eur16]. Further efforts include the Ethics Guidelines for Trustworthy Artificial Intelligence published in 2019 by the High-Level Expert Group on AI [Gui19]. The latter aims to provide a framework for the development

and deployment of trustworthy AI systems, geared towards AI stakeholders, while ensuring the mitigation of potential negative consequences, including bias. Together with the Organisation for Economic Co-operation and Development (OECD) [OEC19], these guidelines aim to leverage AI methods to help humans to identify their biases, and assist them in making less biased decisions accordingly.

The concept of fairness in machine learning is another attempt at reducing algorithmic biases when it comes to automated decision-making, and therefore increasing equal treatment of similar groups or individuals [CCG⁺22]. Increasing algorithmic fairness has been a topic that attracted a lot of attention in the recent years, with over twenty newly proposed notions of fairness [VR18] and countless implementations [PLL⁺23]. Among popular implementations, there are open-source libraries such as AI Fairness 360 [BDH⁺19], Aequitas [SKH⁺19] and FairLeran [WDE⁺23]. The field grows further with propositions to improve performance of discrimination-free algorithms, such as by Papagiannopoulos et al. [PAKM22], and fully-fledged commercial solutions for responsible AI deployment like Deeploy [DWWP24, Dee25] and Code4Thought [Cod25]. Debates and propositions on implementing fairness are well-ongoing within the industry, for example, on contrasting interpretations within the GDPR itself [Cod20], or ill-suited applications of the ‘four-fifths-rule’ [Fai22, WMC22]¹. This signals a lack of agreement on when the multitude of fairness definitions applies and how they are defined exactly [CCG⁺22]. The paper of Verma [VR18] identifies the most prominent definitions of fairness, while providing clear rationales behind these definitions, and illustrating these using an example dataset, in an attempt to address the aforementioned problem.

For this article, the definitions of fairness, identified in the paper of Verma, will be illustrated and applied to a generated fictitious dataset to give a clear and concise overview of these definitions. The dataset represents a typical hiring scenario where hiring candidates are evaluated by a series of algorithms on a set of features. Its size is limited to 225 data points where each represents a fictitious candidates with certain characteristics and the hiring outcome according to four decision algorithms. The overall goal is, thus, to provide readers with a practical notion of fairness, illustrate the definitions on a realistic case study and showcase the potential contradictions between fairness definitions.

The dataset is released as creative commons, and can be used freely for any purpose. If you use it, please cite this paper. The dataset is available on Kaggle: <https://www.kaggle.com/dataset/ictinstitute/utrecht-fairness-recruitment-dataset>.

2. Literature review

There is a lot of interesting research into fairness metrics that provides valid insights for particular cases and situations. Below is an extensive overview of multiple cases that we are aware of. It shows that the current metrics can be applied in practice but also shows the challenges of meeting required fairness targets.

The default fairness approach in machine learning focuses on constructing an optimal ML model subject to fairness constraints (a “constrained optimization problem”). This is exactly where fairness comes into play, as it aims to achieve equal treatment of similar groups or individuals [CCG⁺22].

Achieving fairness however is not that simple. Although many different notions of fairness have been defined, there are still not necessarily a “right” answer for each and every situation, making these definitions therefore often not considered during the development of ML systems.

To address the aforementioned issue, the paper of Verma[VR18] collects the most prominent definitions of fairness for the algorithmic classification problem. For this problem, Verma explains

¹The selection ratio of a minority group should be at 80% of the selection ratio of the majority group – 29 C.F.R. § 1607.4(D).

the rationale behind the defined notions, while also performing a case-study on the topic of loan granting, in order to demonstrate all the different definitions of fairness mentioned in the paper. [VR18]. Compared to the paper of Verma, this paper will provide more detailed examples, using a more compact dataset with multiple sensitive attributes. It thus better illustrates the possible differences between the main fairness notions.

When comparing metrics, one needs to decide what difference in values is acceptable. In the literature the “four fifths rule” or “80% rule” is a often-mentioned rule of thumb commonly used in fairness assessments for machine learning models (and sometimes as a ‘legal waiver’ for legal obligations) [Cod20]. It relies on measuring demographic parity ratios (group fairness), typically labeling outcomes as fair if subgroup ratios exceed 80% (instead of the difference). However, this practice is frequently inappropriate because fairness must be evaluated within the broader socio-technical context, centered around potential harms rather than purely numerical thresholds [Fai22].

The paper of Zliobaite [Zli15] classifies different fairness definitions into different categories (statistical tests, and absolute, structural and conditional measures). Most fairness notions that will be discussed in this article, are also mentioned by Zliobaite, albeit under another name (e.g. difference of means test instead of balance for positive class). As far as the applicability of the fairness definitions is concerned, the only criterion taken into consideration was the type of variable (i.e. binary, categorical, numerical, etc.).

The paper of Berk [BHJ⁺21] aims to clarify a so-called existing trade-off between both fairness and accuracy, as different kinds of fairness. Berk concludes that it is impossible to satisfy all kinds of fairness simultaneously and states that a major complication that occurs in practice, are the different base rates across different legally protected groups. The paper of Berk therefore mentions a need to consider challenging trade-offs. Noteworthy, is that Berk only listed fairness notions surrounding group fairness that are able to be defined by the use of a confusion matrix. Through the use of this confusion matrix and a criminal risk assessment use-case, the paper of Berk aimed to highlight the relationships and possible differences between the fairness notions.

The paper of Castelnovo [CCG⁺22] aims to highlight important aspects in regards to fairness metrics and its respective relationships, as the paper claims that the different definitions of fairness have not yet been analyzed fully in existing literature, which according to Castelnovo, could have a grave impact on both individuals and populations. The paper aims to provide some more structure and order to the many definitions that currently exist on the topic of fairness.

The paper of Hooghiemstra provides a case-study using real-world models. Hooghiemstra also gives a ranking of the different fairness definitions in terms of which definitions could be best used for future AI models [Hoo22]. The paper of Hooghiemstra acts as a good example on exactly how models can be used in order to measure fairness, while considering several different fairness definitions.

The paper of Hutchinson [HM19] compares past and present definitions of fairness on a number of levels, including the criteria for the fairness definitions, the focus of the respective criteria (for example, in the form of a test or a model), the relationship of fairness to individuals and (sub)groups, and lastly the mathematical method used in order to measure fairness (i.e. regression, classification). It ends with a reflection on future prospects when it comes to lessons that can be learned based on the history of test fairness, and the effect these lessons can have when it comes to the future of fairness in ML.

The paper of Snel [SvO22] also describes a case study on how to detect and correct bias in a specific case. In order to achieve this, Snel demonstrates a algorithm-independent bias correction method, which is aimed to result in better ML predictions, mainly when it comes to the likelihood of default. Snel finds that before bias can be eliminated, it is necessary to first apply bias-correction consecutively, which would allow for the mitigation of the impact the bias would

have on various, overlapping groups.

In the paper of Chouldechova [Cho17] hypothetical use cases to connect particular fairness properties of a Recidivism Prediction Instrument (RPI) to a measure of disparate impact, are considered. Chouldechova presents both theoretical and empirical results to illustrate how disparate impact can arise when making use of a RPI that is claimed to not have any predictive bias at all.

The paper of Mittelstadt [MWR23] scrutinizes the use of leveling-down measures in order to achieve fairness and also further investigates how fairML, used for auditing black-box predictive models, can move beyond mere levelling down methods.

The paper of Wachter [WMR21] addresses so-called 'contextual equality', revolving around a critical gap that exists between legal, technical, and organisational definitions of algorithmic fairness. Wachter identifies a critical incompatibility when it comes to already existing works that have been done on the topic of automated and algorithmic fairness, as well as European notions of discrimination, through conducting an analysis of EU non-discrimination law and jurisprudence of the European Court of Justice (ECJ) and national courts.

The paper of Corbett [CDPF⁺17] aims to mitigate disparities when it comes to the decision-making process of releasing defendants awaiting trial back into their respective community. In order to achieve algorithmic fairness, Corbett has proposed several techniques. With this research, the paper of Corbett aims to maximize public safety while satisfying formal fairness constraints aimed at reducing racial inequality.

The paper of Simoiu [SCDG17] develops a new statistical test of discrimination, namely the so-called threshold test, which is able to mitigate the problem of infra-marginality (two groups having different risk distributions, resulting in different decision-making across groups, even without discrimination). Simoiu does this estimating both the risk distributions and the most appropriate decision thresholds. These two aspects are then applied to be tested against a data set of 4.5 million police stops, present in North Carolina.

The paper of Galhotra [GBM17] gives definition to software fairness and discrimination and develops a method that would allow the measuring of software and whether it discriminates, and if so, how much it actually discriminates. This is done with the aim of identifying possible causes for discriminatory behavior. Galhotra mentions their approach Themis, which helps with measuring discrimination, uses Themis for 20 different software systems. The paper of Galhotra overall demonstrates that fairness testing is crucial when it comes to software development, especially in domains where possible discrimination is prevalent, and provides possible tools that can be used when trying to measure software discrimination.

The paper of Prince [PS19] offers various potential strategies when it comes to combating the risk of proxy discrimination by AIs. Proxy discrimination is described by Prince as a particularly pernicious subset of disparate impact, which is able to harm individuals that belong to a protected class. The paper of Prince further argues that AI and big data play a crucial role when it comes to the aforementioned issue of proxy discrimination.

The paper of Wisniewski [WB21] introduces an R package fairmodels, which can be used to solve the issue of existing bias in classification models, and can help validate fairness in a more efficient and flexible manner. The package revolves around all sorts of methods surrounding the mitigation of biases and therefore aims to help diminish discrimination in multiple classification models.

The paper of Zafar [ZVRG17] introduces a flexible mechanism that is able to design fair classifiers, using a new measure of decision boundary (un)fairness. This mechanism is expressed through the use of two classifiers, namely Support Vector Machines (SVM) and Logistic Regression. Using these classifiers, the paper of Zafar is able to show using real-world data, which usually only costs very little in terms of accuracy, that the mechanism allows to have precise control on fairness, more specifically the degree in which it is achieved.

The paper of Dwork [DHP⁺12] introduces a framework for fair classification that is able to compromise on ensuring that people are treated fairly and making use of a task-specific metric that is able to assess how similar individuals are, while keeping the specific classification task in mind.

Finally, the work by Papagiannopoulos et al. [PAKM22] presents and validates a data science approach to improve fairness metrics while retaining an algorithm's predictive performance. Many previous methods to mitigate bias and discrimination in machine learning fall short on accuracy performance. Papagiannopoulos et al.'s approach implements a fairness-aware bagging procedure leveraging the randomness in resampling methods to help improve models' accuracy.

3. Fairness definitions

3.1 Types of machine learning tasks

Machine learning tasks are often classified according to the type of output required, for instance in textbooks[Ger19]. The most common groups of tasks are *regression* (predicting a continuous value, such as value of an item or next day temperature), *classification* (assigning an item to the correct class out of a finite number of classes) and *binary classification* (classification with only two classes, a positive/selected class and a negative class). Knowing the class of a task is important to make sure you apply the right metrics: metrics like accuracy are suitable for classification but not for regression. Precision and recall are defined for binary classification, but not for other tasks.

Similarly, many fairness metrics are only suitable for specific tasks. In this paper we therefore introduce the definitions for identifying groups of tasks that require different treatment in fairness.

- A task is *fairness-sensitive* if individual decisions have a significant positive or negative impact on people or groups of people. The people affected are often not the users of AI systems. Depending on the context, they can be candidates, buyers, owners, citizens, residents, students or employees.
- We propose to call the people who are impacted by the decisions in a fairness-sensitive task the *impacted individuals* or informally *individuals*. In this paper we will also use the term *candidates* since this is the preferred term when in case of hiring decisions.
- A *desired-selection task* is a binary classification tasks where individuals would like to be selected. Examples are hiring processes where people would like to be hired, university acceptance processes, or loan approval processes where people would like to be offered the loan. The example dataset introduced in this paper presents a desired-selection-task.
- A *desired-accuracy task* is a classification task where individuals just want to be classified correctly. An example of such task is the case where a machine learning algorithm is used to estimate clothing sizes, and individuals just want to receive the right size.
- A *desired-high-score task* is a regression task where individuals prefer to get a higher score over any lower score. An example of such task would be a fitness test, essay grading task or any assessment.
- A *desired-low-score task* is a regression task where individuals prefer to get a lower score over any higher score. An example of such task would be credit scoring or risk estimation, such as [SvO22].
- A *undesired-selection tasks* is a binary classification tasks where individuals would like to not be selected. An example is a tasks were people are selected for rejection, or where transactions by individuals are selected as fraudulent.

Note that any *undesired-selection task* can be converted into a desired-selection tasks by changing the labels: instead of saying an individual is selected for rejection or identified as fraudulent (the undesired outcome), one can state that the individual is not selected for the desired outcome (immediate approval or no-fraud). We recommend this conversion and will assume tasks are not *undesired-selection tasks*.

Note also that in many tasks it does not have to be the case that all individuals have the same preferences over outcomes. For instance in the case of estimating house values, some house owners might prefer a lower value since this may lead to lower taxes, some house owners might prefer a higher value since they intend to sell their house and a higher value makes the house more attractive to buyers, and other buyers prefer a predicted value close to the real value to help with financial planning.

3.2 Basic concepts extended to groups

Before diving into the definitions of fairness metrics, we first introduce a set of formal definitions of basic mathematical concepts used in many fairness definitions. These are listed in Table 1. Table 1 defines several basic concepts often used in machine learning. These definitions are commonly used to measure the performance of decision algorithms on a complete data set. We redefine these concepts so that they are defined for any subgroup g of cases (candidates for our dataset). Since the task of the algorithm is to predict a certain property, each case in the data set either has the property (a positive case) or does not have the property (a negative case). An algorithm makes a prediction for each case and the prediction (predictive class) can be compared to the true answer (actual class).

In the literature of fairness metrics, metrics are often compared without specifying whether an exact equality is required or not. This is a useful simplification in a theoretical discussion. However, in many real situations an exact equality is not be attainable and is also not required. Therefore, one must define *approximate equality*. We define that two values are *approximately equal* when the difference of the two values is lower or equal than a threshold of 0.1 ($th = 0.1$) to provide a clear and practical definition of what we consider fair. Algorithmic fairness assessors and auditors can also use this definition by referring to this paper, or should explicitly define a different threshold. When given metric values for different groups are approximately equal, we consider that fairness holds between those two groups according to that metric.

Table 1. Definitions of performance metrics for evaluating binary classification algorithms on an overall dataset.

Name	Acronym	Description
True Positive	TP(g)	The number of true positive cases in a subgroup g is the number of cases in g that are actually positive and selected.
True Negative	TN(g)	The number of true negative cases in a subgroup g is the number of cases in g that are actually negative and not selected.
False Positive	FP(g)	The number of false positive cases in a subgroup g is the number of cases in g that are actually negative and selected
False Negative	FN(g)	The number of false negative cases in a subgroup g is the number of cases in g that are actually positive and not selected.
All Positives	AP(g)	The total number of positive cases in a subgroup g is the total number of cases in g that are selected.
All Negative	AN(g)	The total number of negative cases in a subgroup g is the total number of cases in g that are not selected.
Cardinality	C(g)	The number of elements/cases in the set of subgroup g .
Exact equality	$a = b$	Two numbers a and b are exactly equal if their difference is zero. Thus $0.4000 \cong 0.4000$ but not $0.3998 \cong 0.4000$.
Approximate equality	$a \cong b$	Two numbers a and b are approximately equal if their absolute difference is 0.1 or less. Thus $0.41 \cong 0.45$ but not $0.39 \cong 0.51$.

3.3 Individual fairness definitions

Individual fairness means that classification algorithms should predict similar outcomes for similar individuals in a given context. That is, two individuals with the same or similar attributes

should lead to the same prediction [DHP⁺12]. Individual fairness definitions typically include multiple sensitive attributes (like age and gender in our dataset). They do not look solely at gender or reduce age to two groups. It is often implemented using a similarity metric or a distance function [DHP⁺12].

Individual fairness has been criticized on the basis that individual fairness approaches do not guarantee fairness and often include existing biases, for instance in the definition of the similarity metric. This is described by Fleisher[Fle21]. It has also been discussed that individual fairness and group fairness could lead to apparent conflicts, e.g. when a male candidate is not invited for an interview while he has similar test scores as some female candidates that were invited, he could argue that this violates individual fairness. A detailed philosophical discussion of this apparent conflict is given by Binn[Bin19].

In our view, individual fairness is an important principle but difficult to apply in practice. First of all it is not easy to apply since one needs to decide what similarity metric to use: there is no obvious similarity metric that includes age, gender ethnicity and other sensitive features in a natural way. Secondly, the use of metrics instead of groups makes it much harder to set explicit acceptance criteria, and thus makes audits and assessments more difficult in practice. In this paper we therefore do not include individual fairness based in similarity metrics. Our approach, inspired by the recommendations of Buolamwini and Gebru[BG18] is to use group fairness definitions but with multiple groups instead of just one group. If you use this approach and combine multiple sensitive features to define small subgroups (e.g. darkskinned women for facial recognition[BG18] or older female candidates you can also make sure the risks towards all individuals are considered.

We do agree with the principle behind individual fairness based on similarity metrics, that one should consider the individual situation and consequences for individual candidates when discussing fairness. As Fleisher[Fle21] points out, the technical individual fairness definitions are not sufficient to solve this problem. We recommend that organisations look beyond metrics and consider each individual decision carefully. Metrics can help find out what the negative decisions are that candidates could consider unfair. Following Table 1, recall can be considered as a preferred metric for measuring individual fairness towards qualified candidates, since it is based on avoiding false negatives (the lower FN, the higher recall): $\text{Recall} = \frac{TP}{TP+FN} = \frac{TP}{AP}$.

One could also argue that the overall selection rate (Hiring rate = $\frac{AP}{C}$) is the best metric to consider, since this is based on hiring the most candidates and thus satisfying the needs of most candidates. Hiring all candidates however is not a practical for most businesses.

3.4 Group fairness definitions

The idea behind group fairness metrics is that there are specific groups of individuals that might be discriminated against, and that you compare metrics for your algorithm on these groups against the same metrics for the entire dataset. Group fairness criteria are applied at group level where *sensitive* attributes define a group (e.g., age group, gender, sexual orientation, etc.). The problem with group fairness is that there is not one obvious metric (e.g. accuracy) to consider. As we will show, there are many metrics that one could consider. Considering the overview of previous literature, we follow along the fairness definitions of Verma and Rubin [VR18] as their work summarizes the most prominent definitions of fairness for algorithmic classification problems. Whereas their paper goes over 20 fairness metrics, we focus on common metrics to provide a practical, easy to follow demonstration: Demographic parity (statistical parity), Conditional statistical parity, Predictive Parity (PPV-parity), Predictive equality (FPR-parity), Equality of Odds, Treatment Equality, Test fairness, Well-calibration, Fairness through unawareness and Fair inference.

The selected fairness definitions collected by Verma [VR18] are presented in Table 2. All definitions refer to a specific group to be protected (e.g., women, candidates over a certain age, immigrant workers etc.). We use p for the protected group and μ for the set of all people outside the

protected group. In our examples, female candidates are the protected group. In practical situations one must consult domain experts who understand the legal and ethical considerations of the case in order to determine what the to be protected groups are.

Table 2. Selected fairness metrics definitions from Verma [VR18], based on algorithm performance metrics of Table 1.

Name	Definition	Description
Group fairness [DHP ⁺ 12] (benchmarking [SCDG17], equal acceptance [Zli15], statistical parity [DHP ⁺ 12])	$PPP(u) \cong PPP(p)$ where $PPP(g) = \frac{AP(g)}{C(g)}$	This definition is satisfied in case individuals in both protected (p) and unprotected (u) groups have an equal probability of being selected by the classifier: Positive Predicted Probability (PPP), or being assigned to the positive predicted class.
Conditional statistical parity [CDPF ⁺ 17]	$PPP(u, l) \cong PPP(p, l)$ where $l \in L$	This definition is satisfied in case individuals in both protected (p) and unprotected (u) groups have an equal probability of being selected by the classifier (PPP) given one or more legitimate attributes (L) such as, experience, place of residence, etc.
Predictive parity [Cho17]	$PPV(u) \cong PPV(p)$ where $PPV(g) = \frac{TP(g)}{AP(g)}$	This definition is satisfied in case the Positive Predicted Value (PPV) – or precision – of both the protected and unprotected groups is equal, meaning that the probability of an applicant that is predicted to be selected by the classifier, is equal to actually being hired.
False positive error rate balance [Cho17] (predictive equality [CDPF ⁺ 17])	$FPR(u) \cong FPR(p)$ where $FPR(g) = \frac{FP(g)}{N(g)}$	This definition is satisfied in case the False Positive Rate (FPR) for both the protected and unprotected is equal, i.e, the probability of an applicant that is actually not hired for a company, to be incorrectly selected by the classifier, is equal for both groups.
Equalized odds [HPS16]	$(TPR(u) \cong TPR(p)) \wedge (FPR(u) \cong FPR(p))$ where $TPR(g) = \frac{TP(g)}{P(g)}$	This definition is satisfied in case the True Positive Rate (TPR) for both the protected and unprotected groups are equal, meaning that the probability of an applicant that is actually hired for a company to be selected by the classifier and the probability of an applicant that is actually not hired to not be selected by the classifier, is equal for both groups.
Treatment equality [BHF ⁺ 21]	$\frac{FN(u)}{FP(u)} \cong \frac{FN(p)}{FP(p)}$	This definition is satisfied in case the ratio of False Positive (FP) to False Negative (FN) is equal for both protected and unprotected groups.
Fairness through unawareness [KLRS17]	Test method: classification output equals to classification output with swapped sensitive attribute	This definition is satisfied in case no sensitive attributes are used in the decision-making process, meaning that gender-related or nationality-related features are not used when training the classifier, so whether an applicant is hired for a company or not, cannot rely on these features. This also means that the classification outcome should be the same for any applicant that have the same attributes. It is a similarity measure not relying on sensitive attributes, with no mathematical definition.
Fair inference [NS18]	Test method: no illegitimate paths to prediction in a causal graph.	This definition is satisfied in case paths in causal graphs are classified as either legitimate or illegitimate. In case there are no illegitimate paths from A to Y , a causal graph can be considered to satisfy the notion of fair interference. No mathematical definition, as it is based on causal reasoning, assuming a given causal graph.

4. Dataset overview

To illustrate how fairness metrics can be employed in practice, we generated a dummy dataset of 225 candidates applying for a fictitious job position as a security guard. Each candidate is described by age, gender and whether they live close to the job location. The job requires candidates to be able to pass a physical strength and speed test to qualify for the position, therefore a variable encodes their test result. The employer estimates a business value for a successful hire of suitable candidate (a not suitable candidate has a negative value). Similarly, each candidate has

an estimated loss if not hired (a skilled candidate loses more with respect to an unskilled candidate given, e.g., previous investments in training). Finally, each candidate is evaluated against five different classification methods and each method's hiring decision is recorded. Specifically, each candidate record consists of the following variables: `mainid`, `age`, `gender`, `testresult`, `livesnear`, `suitability`, `Value-hired`, `Candidate-loss-nothired`, `should-hire`, `hired-by-expert`, `A1`, `A2`, `A3` and `A4`. The explanation of each variable follows in Table 3. Appendix Appendix 1 presents additional details on the dataset.

Table 3. Variables description.

Variable	Description	Count	Example
<code>mainid(hidden)</code>	Unique identifier	225	225153
<code>Age(feature, sensitive)</code>	Age, sensitive feature (20-49)	30	25
<code>Gender(feature, sensitive)</code>	Gender, sensitive feature (female, male)	2	female
<code>testresult(feature)</code>	Outcome of strength and speed tests as job requirement (0-2)	5	1.5
<code>livesnear(feature)</code>	Does the candidate live near to job location?	2	0
<code>Suitability(target)</code>	Candidate suitability as function of testresult (1-4)	4	3
<code>Value-hired</code>	Business value when hired (-10000-10000). Unsuitable candidates have a negative business value.	4	5000
<code>Candidate-loss-nothired</code>	Candidate loss when not hired (10 - 5000). The loss is higher for more suitable candidates. It is positive for each candidate since all candidates want to be hired	3	5000
<code>Should-hire(target)</code>	Target variable for classification methods. It is derived from <code>value-hired</code> , which is the ground truth	2	1
<code>hired-by-expert</code>	Outcome classification method by human expert (discriminates on <code>Suitability</code> and <code>livesnear</code> with additional noise)	2	0
<code>A1(testresult)</code>	Outcome classification method (uses <code>testresult</code>)	2	1
<code>A2(testresult,30under)</code>	Outcome classification method (uses <code>testresults</code> and <code>Age</code>)	2	1
<code>A3(Age,Gender, test)</code>	Outcome classification method (uses <code>testresults</code> , <code>Age</code> and <code>Gender</code>)	2	1
<code>A4(positive-dicr)</code>	Outcome classification method (uses <code>testresults</code> and <code>age</code> and positively discrimination on <code>Gender</code>)	2	1

5. Fairness results

5.1 Recruitment as an optimization problem

Our scenario for illustrating fairness metrics usage consists of a fictitious physical security company that wants to fulfill their recruitment needs while maximizing the business value of each hire. On the other hand, each candidate wants to be hired, however certain candidates are more or less suitable and, consequently, would incur in a loss on their own if not hired. We codified such considerations as `Value-hired` and `Candidate-loss-nothired` in the dataset (see Table 3). It is thus evident that the employer seeks to hire candidates that provide the best business value, and at the same time minimize the candidate loss by selecting the right candidates for the job. However, the incurred loss for the business and a candidate is asymmetrical: the loss for the business is high for a wrong hire (e.g., 10000) as opposed to a loss of an unqualified candidate (e.g., 10). In other words, a false negative is expensive for the candidate, and both false positives and false negatives are expensive for the business.

Following Table 3, each candidate in the hiring dataset has a `Should-hire` label, 1 or 0, that means whether the candidate should be hired or not. This label represents the ground truth,

i.e., the ideal result of the hiring procedure. Each classification algorithm (hired-by-expert, A1, A2, A3 and A4) outputs a hiring decision based on factors described in Table 3. For example, A2 outputs the hiring decision based on a threshold for the physical test result and age limit. Hired-by-expert, on the other hand, simulates a human making the hiring decision based on a subjective suitability scale (Suitability in Table 3), the candidate residence location and added noise (to simulate human errors). We can, therefore, evaluate the performance of each classification method based on its error rate against the ground truth.

Table 4. Overall performance metrics of the 5 classification methods.

Method	TP	TN	FP	FN	Precision	Accuracy	Recall	F1 Score	Value	Candidate-loss
hired-by-expert	54	141	9	21	0.86	0.87	0.72	0.78	295,010	129,180
A1(testresult)	57	110	40	18	0.59	0.74	0.76	0.66	95,120	102,980
A2(testresult,30under)	49	127	23	26	0.68	0.78	0.65	0.67	205,100	145,130
A3(Age,Gender, test)	53	127	23	22	0.70	0.80	0.71	0.70	230,090	126,120
A4(postive-dicr)	33	122	28	42	0.54	0.69	0.44	0.49	10,080	227,060

Table 4 shows the basic performance metrics for each classification algorithm. The hired-by-expert has the overall best performance, e.g., few FPs (9) and FNs (21), almost on par with A3. A1 and A4 have the worst performance of them all, with a high amount of FPs (40) and FNs (42), respectively. We included two extra columns for business value and candidate loss that represent the total value for the right (TP) and wrong decisions (FP). For instance, the algorithms A2 and A3 have a comparable performance while providing significantly different values for hire due to the underlying features used by the classifier. We can visualize it better in Figure 1.

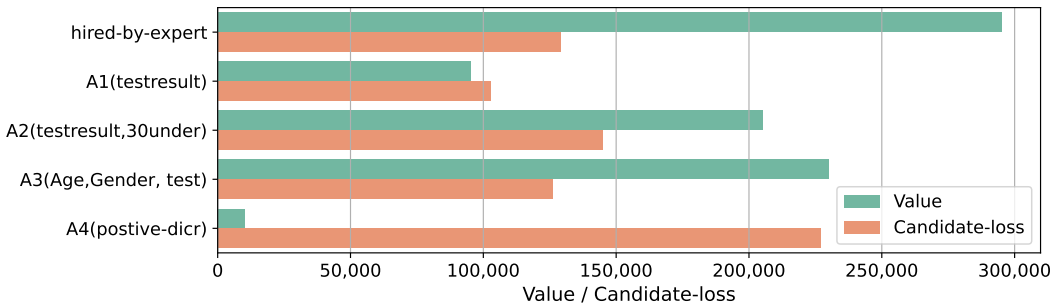


Figure 1. Barchart of business value and candidate loss. The differences between the classification algorithms are evident in terms of monetary value/losses, with hired-by-expert scoring best and A4 scoring the worst.

Figure 1 suggests that the business value is maximized by the hired-by-expert with A3 following suit. Interestingly, this is not necessarily true for the candidates point of view: the lowest overall candidate loss happens for A1. The lowest overall value, for both the employer and the candidates, is A4. We may conclude from this evaluation that hiring expert and algorithm A3 are the best options for a business decision.

To appreciate the differences of how each algorithm classifies candidates to hire with the related discrimination conditions, we show a scatterplot of hiring predictions by age and test result for the actual hires, hired by expert and A1 in Figure 2. Furthermore, Figure 2 encodes the marker color for the hiring decision (algorithms output) and marker shape for the gender variable. For example, the figure shows that A1(testresult) advises to hire only individuals that pass the test with a score higher than 1, irrespective of age and gender. On the other hand, the hiring by expert does not rely exclusively on the test score and includes candidates even with lower test scores. The

ground truth (Should-hire), displays the actual hires across various ages, gender and test results. Figure 6 in Appendix 1 shows hiring decisions for the remaining algorithms.



Figure 2. Scatterplot of hiring outcomes for ground truth (1st plot), hiring-by-expert and A1 methods against test score and age. The markers are coded by shape (female/male) and color (hired/not hired). The differences in hiring decision are evident by looking at the three methods discriminate by test result, age and gender. A jitter is added to the Y-axis to improve the visualization as otherwise the markers would be superimposed on discrete values [Lee17].

It is therefore clear that the various algorithms of rely on different features to label a candidates as hired/not hired, which conditions they overall performance. What is interesting to explore is how do these classification methods, including the hired-by-expert, fare in terms of fairness.

5.2 Individual fairness

In Section 3.3, we mentioned that we would not consider individual fairness in the normal way using similarity metrics. Instead we consider the fairness towards each individual using the metrics hiring rate and recall. These metrics capture how many candidates get the desired outcome. We will also look at precision to consider the quality of the algorithms from the employer point of view. Figure 3 shows the metric values for each algorithm.

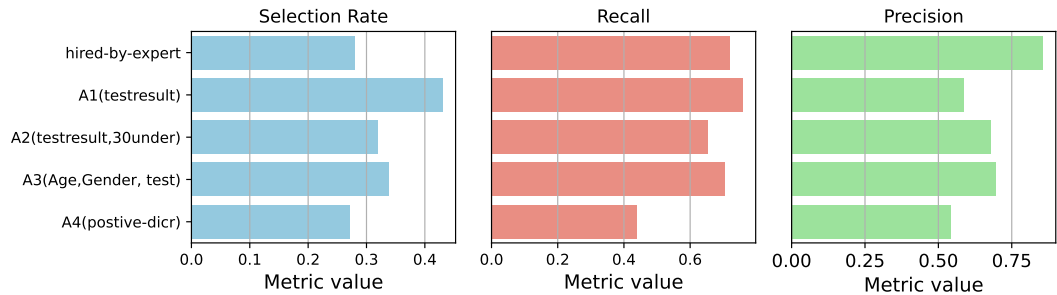


Figure 3. Selection rate, Recall (candidate perspective) and Precision (business perspective) across algorithms.

It is immediately clear that the algorithms differ significantly on these overall metrics. The highest selection rate (hiring rate) belongs to A1 which would hire more than 40% of candidates. Recall, on the other hand, captures the fraction of relevant candidates being hired, with A1 and hired-by-expert being the most fair to the individual. Finally, the precision metric tells us that hired-by-expert is by far the best one minimizing false positives (costly errors for the employers).

From the candidate perspective, we may conclude that the preferred algorithms by most candidates in terms of individual fairness is A1, with the highest Recall and the higher likelihood of being hired (an overall large selection ratio). From the business perspective, hired-by-expert is the

best option as precision is maximized. Candidates have a lower chance of getting hired (low selection ratio). There is thus an apparent conflict between the needs of the business and the needs of candidates. In a detailed fairness assessment, one could decide trying to finetune all algorithms to have the same recall before comparing these algorithms. This would make the algorithms easier to compare. Such finetuning is however not always possible: in many cases including with this dataset there are not enough different test scores to get every selection rate.

These observation about the overall fairness towards individuals, however, do not tells us anything about specific groups that may be discriminated (as suggested by Figure 2). To say whether our selection algorithms are fair towards different groups in our candidate pool, we need to investigate group fairness.

5.3 Group fairness

Following the overview of the definitions in Table 2, it is possible to illustrate the fairness metrics using our hiring dataset. In the dataset, there are two sensitive variables: `Age(feature, sensitive)` and `Gender(feature, sensitive)`. An additional variable that can be used to discriminate candidates is `livesnear(feature)`. For our examples, we need to select the protected group(s) within the sensitive attributes. A sensible choice that accounts for our realistic scenario is candidates over 40 years old ($Age > 40$) and female candidates, as both groups can be potentially disadvantaged (legitimately or not). One could decide based on domain knowledge or based on expert interviews tp also consider ‘lives near’ as a sensitive attribute. One could also argue that this is non-sensitive, to differentiate candidates by a legitimate attribute. In the following tables, we compute the first five fairness metrics to allow a tabular representation for readability reasons. The remaining two (Fairness through unawareness and Fair inference) are described in their own subsections.

In Table 5, we compute the defined fairness metrics by accounting for the sensitive attribute Age with $p = Age > 40$ (and $u = Age \leq 40$). We can immediately see that Group fairness is not satisfied for all algorithms except A4. Specifically, the selection rate for candidates over 40 is consistently lower than for individuals below 40. A2 and A3 do not select individuals > 40 at all, which has the consequence of failing all the fairness tests in Table 5. Predictive Parity and Treatment equality are not satisfied for all the classification methods. Predictive equality is the satisfied for all cases except A2 and A3. Overall we can conclude that all algorithms are mostly unfair towards candidates over 40, with few exceptions for hired-by-expert and A4.

Table 5. Fairness metrics for sensitive attribute Age group ($p = Age > 40$ and $u = Age \leq 40$).

Metric	hired-by-expert			A1(testresult)			A2(test,30under)			A3(Age,Gender, test)			A4(postive-dicr)		
	>40	≤40	Fair	>40	≤40	Fair	>40	≤40	Fair	>40	≤40	Fair	>40	≤40	Fair
Group fairness	0.35	0.15	X	0.51	0.28	X	0.49	0.00	X	0.52	0.00	X	0.27	0.28	✓
Pred. par. (PPV)	0.88	0.75	X	0.67	0.32	X	0.68	0.00	X	0.70	0.00	X	0.67	0.32	X
Pred. eq. (FPR)	0.07	0.05	✓	0.30	0.23	✓	0.27	0.00	X	0.27	0.00	X	0.15	0.23	✓
Eq. odds (TPR,FPR)	0.71, 0.75	0.07, 0.05	✓	0.79, 0.58	0.30, 0.23	X	0.78, 0.00	0.27, 0.00	X	0.84, 0.00	0.27, 0.00	X	0.41, 0.58	0.15, 0.23	X
Treat. eq. (FN/FP)	3.00	1.00	X	0.52	0.33	X	0.61	inf	X	0.43	inf	X	2.85	0.33	X
Count	78	147		78	147		78	147		78	147		78	147	

Table 6 shows gender as the sensitive attribute, with female candidates being the protected group (p =female). Group fairness is achieved only in A2 and A4, with A2 not making strong distinctions between female and male candidates. Predictive parity is satisfied for all except A3 (more precise for female candidates), and Predictive equality is not true for A1 and A3 (unbalance of FPs). Interestingly, only the ‘human expert’ method and A3 are fair by Equality of odds, which is one of the strictest fairness metrics. Treatment equality is not satisfied by any of the methods. With sensitive attribute Gender, three algorithms (hired-by-human, A2 and A4) can be considered relatively fair by three different metrics. It appears, therefore, that the classification methods are overall more fair towards female candidates, as opposed to candidates over 40 years old.

Table 6. Fairness metrics for protected attribute Gender (p =female and u =male).

Metric	hired-by-expert			A1(testresult)			A2(test,30under)			A3(Age,Gender, test)			A4(postive-dicr)		
	female	male	Fair	female	male	Fair	female	male	Fair	female	male	Fair	female	male	Fair
Group fairness	0.34	0.16	X	0.48	0.33	X	0.35	0.27	✓	0.43	0.16	X	0.24	0.33	✓
Pred. par. (PPV)	0.84	0.92	✓	0.61	0.52	✓	0.73	0.55	X	0.70	0.67	✓	0.56	0.52	✓
Pred. eq. (FPR)	0.09	0.02	✓	0.31	0.20	X	0.16	0.15	✓	0.21	0.07	X	0.18	0.20	✓
Eq. odds (TPR,FPR)	0.73, 0.72	0.02, 0.09	✓	0.87, 0.73	0.20, 0.31	X	0.73, 0.63	0.15, 0.16	✓	0.53, 0.75	0.07, 0.21	X	0.87, 0.33	0.20, 0.18	X
Treat. eq. (FN/FP)	4.00	2.12	X	0.17	0.57	X	0.44	1.57	X	1.75	0.79	X	0.17	2.50	X
Count	75	150		75	150		75	150		75	150		75	150	

Finally, let’s consider the case of ‘livesnear’ attribute. Table 7 shows the fairness metrics with sensitive attribute Gender for candidates that do not live near the job location (livesnear=0). By comparing Tables 7 and 6, it is clear that certain fairness definitions for the protected group p =female do not hold anymore when considering the control variable livesnear. Specifically, for livesnear=0, Group fairness for A2 is lost (here called Conditional Group fairness as per definitions in Table 2). The same for Predictive parity, according to which, fairness was achieved in Table 6 for the majority of the classification methods, but in Table 7 it holds only for A4. Interestingly, Predictive equality (similar FP rates) does not change for Gender attribute with and without considering lives near.

There are other possible combinations to check for fairness, such as using ‘livesnear’ as a sensitive feature, female candidates over 40 (p =female and Age>40) or even female candidates over 40 living far away (p =female and Age>40 and $l=0$). However, we omit the discussion of such cases for readability reasons and, instead, provide an example in Appendix 1 (where the p =female and Age>40 combination in Table 11 leads to only Predictive equality to be satisfied with hired-by-expert).

5.3.1 Fairness through unawareness

Another interesting fairness definition in Table 2 is Fairness through unawareness whereby the algorithms should not use any sensitive feature for classification (Age and Gender). This is an important fundamental fairness definition that can be required in certain domains. E.g. Article 9 of the GDPR[Eur16] explicitly forbids the processing of the special features ethnic origin, political opinions, religious or philosophical beliefs, trade union membership or sexual orientation unless there is a strong reason to do so. So any algorithm using these features in a situation like hiring is illegal in the EU. Many national laws have similar anti-discrimination clauses. Fairness through

Table 7. Fairness metrics for attribute Gender and control attribute ‘livesnear’ (p =female, μ =male and l =livesnear).

Metric	hired-by-expert			A1(testresult)			A2(test,30under)			A3(Age,Gender,test)			A4(postive-dicr)		
	female	male	Fair	female	male	Fair	female	male	Fair	female	male	Fair	female	male	Fair
Conditional Group fairness															
livesnear=0	0.05	0.19	X	0.24	0.50	X	0.22	0.40	X	0.14	0.47	X	0.24	0.20	✓
Pred. parity (PPV)															
livesnear=0	0.50	0.69	X	0.44	0.60	X	0.50	0.71	X	0.80	0.64	X	0.44	0.50	✓
Pred. equality (FPR)															
livesnear=0	0.03	0.09	✓	0.16	0.32	X	0.12	0.18	✓	0.03	0.27	X	0.16	0.16	✓
Equality of odds (TPR,FPR)															
livesnear=0	0.20, 0.35	0.35, 0.09	X	0.80, 0.16	0.81, 0.32	X	0.80, 0.12	0.70, 0.18	✓	0.80, 0.03	0.81, 0.27	X	0.80, 0.16	0.27, 0.16	X
Treat. eq. (FN/FP)															
livesnear=0	4.00	4.25	X	0.20	0.36	X	0.25	0.75	X	1.00	0.42	X	0.20	2.71	X
Count															
livesnear=0	37	70		37	70		37	70		37	70		37	70	

unawareness through unawareness can be evaluated for algorithms if one has access in two ways.

- One can inspect the code to understand the underlying logic, and check if the algorithm makes use of sensitive features. This can for instance be done to see that A1 does not use gender or age in its decision making. One could call this a white-box approach since you would look into the algorithm itself.
- One can run the algorithm on modified data. You change the values of all sensitive features, to see if there is any combination of sensitive feature values and see if there is any change in outcome. If the outcome changes for any input and any combination of changed features, the algorithm does not satisfy fairness through unawareness. This method is the black box approach since you consider the algorithm a black box that you cannot inspect.

We used the black-box approach to determine whether the algorithms A1,A2,A3 and A4 satisfy fairness through unawareness with respect to Age and Gender. We could not determine whether the hired-by-expert method satisfies fairness through unawareness. If one does not have access to the code or rules used, one cannot determine if a method satisfies fairness through unawareness. This is an important practical limitation for this fairness definition.

One imperfect but practical way to report on whether an algorithm satisfies fairness through unawareness is by comparing the methods’ performance metrics with modified data with those of the original dataset (see Table 4): in case an algorithm differs in terms of TPs and TNs (or FPs and FNs), then they can be considered as not satisfying the Fairness through unawareness definition.

Table 8. Performance metrics compared with Age and Gender unaware classification methods.

Method	Original				Age Unaware				Gender Unaware			
	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
A1(testresult)	57	110	40	18	57	110	40	18	57	110	40	18
A2(testresult,30under)	49	127	23	26	57	110	40	18	49	127	23	26
A3(Age,Gender, test)	53	127	23	22	73	69	81	2	58	116	34	17
A4(postive-dicr)	33	122	28	42	27	136	14	48	26	131	19	49

Table 8 shows that the hired-by-expert, A3 and A4 change the classification outcome for both sensitive features and, therefore, do not satisfy the definition of fairness through unawareness.

5.3.2 Fair inference

The definition of fair inference assumes that there is a known causal graph that shows the relation between features in the dataset. This is often not the case: one needs additional information from domain expert interviews or from inspecting the code of algorithms to obtain such graph. In many real world situations, the relations between features are not known and a causal graph cannot be made or can only be made by making assumptions. This is an important practical problem of inference based fairness metrics.

For our dataset, we do know have this information and we can create a directed graph with features as nodes relationships between features as edges. The solid lines are relations derived from inspecting the algorithm. The dotted lines are relations based on claims from domain experts and are thus somewhat subjective. Figure 4 shows A1, A3 and hired-by-expert as an example of casual graph. The ‘hire’ node is the outcome of the classification method, the remaining nodes are attributes of the dataset, the edges represent the relationship (a dotted edge represents a relationship with a proxy variable [VR18]). Age and Gender are our sensitive attributes.

Fairness in Fair inference has been summarized in secondary literature as the absence of an illegitimate path from any attribute to the outcome node[VR18]. In the original source[NS18] they use a rather complicated statistics based approach using probability distributions and tolerances to formalize this notion. This makes this fairness definition hard to apply.

To illustrate the principle behind fair inference, we made an example for the method A1(testresult) in Figure 4. A1 uses the ‘Test result’ to make the hiring decision, and ‘Test result’ derives from a ‘Speed test’ and a ‘Lift test’ (see Table 3). Age and Gender have an impact the latter two (the domain knowledge suggests that age affects physical speed and gender strength). There is a path between the sensitive variables and the outcome, consisting of two edges. As another example, the directed graph of A3(Age,Gender,test) has a direct path from Age and Gender attributes that condition the ‘hire’ node and is thus definitely not fair according to fair inference. Finally, we tried to make a direct graph for hired-by-expert based on assumed domain knowledge and observable correlations in the dataset. Suitability derives directly from ‘Speed’ and ‘Lift test’ (each with a sensitive proxy variable).

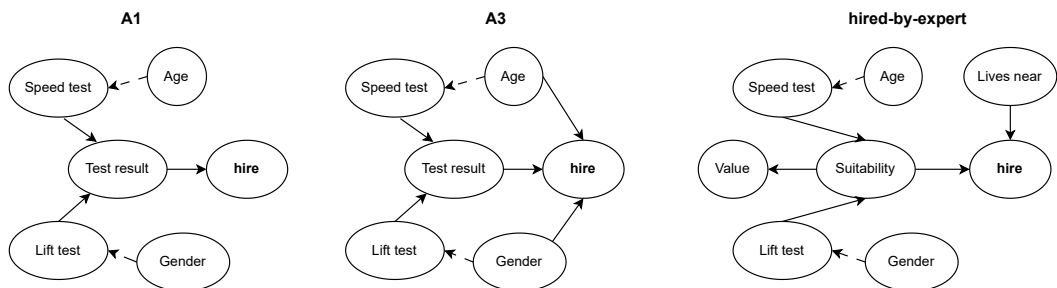


Figure 4. Causal graphs for A1, A3 and hired-by-expert algorithms based on the dummy dataset generation. Nodes represent attributes (‘hire’ is the outcome variable) and edges relationships (dotted edge connects a proxy variable).

6. Discussion and Conclusion

6.1 Discussion

In this paper we have shown the complexity of assessing fairness.

First of all, through illustrating the various fairness definitions identified by the paper of Verma [VR18] on our example dataset and multiple algorithms, we proved that different fairness definitions

are indeed different and will lead to different outcomes. One example is that we show the possibility of indirect discrimination: that it is possible for algorithms to be unfair even when the algorithm does not discriminate directly. By looking at the various fairness tests in Section 5.3, it becomes evident that a gender bias against female applicants only occurs for some fairness definitions and some algorithms. For instance, hired-by-expert and A2 are fair algorithms according to Group fairness (A2), Predictive parity (hired-by-expert), Predictive equality and Equality of odds (Table 6) and Gender unawareness (A2 in Table 8). Age-wise, the majority of algorithms score low on fairness and thus discriminate more strongly for candidates over 40 (Table 5). Of course, A4 also scores well in fairness for both Gender and Age as it is designed to include female and >30 candidates.

What we would like to focus on, however, are the contradictions between the various metrics. A better way to visualize such contradictions is in Table 9. Table 9 show that four different algo-

Table 9. Examples of contrasting fairness outcomes across algorithms and fairness definitions.

	hired-by-expert	A1(testresult)	A2(test,30under)	A3(Age,Gender, test)	A4(positive-dicr)
Group fairn. - Age	X	X	X	X	✓
Eq. odds - Age	✓	X	X	X	X
Fairn. unawar. - Age	unknown	✓	X	X	X
Group fairn. - Gender	X	X	✓	X	✓
Pred. parity - Gender	✓	✓	X	✓	✓
Fairn. unawar. - Gender	unknown	✓	✓	X	X

gorithms can be considered fair for candidates >40 and unfair at the same time according to three different fairness definitions (Group fairness, Equality of odds and Fairness unawareness). Similarly, Gender-wise, Group fairness holds for A2, while Predictive parity holds for all *except* A2. Fair inference flags any of classification methods that directly employ Age or Gender as unfair.

This brings us to the crucial point where there is no simple rule on which is the best fairness definition to use in all situations and, sometimes, not even in a given situation such as ours.

Secondly we have shown that importance of specifying which features are considered sensitive. It is often not the case that there is only one sensitive feature. In this dataset there are two features that are often considered sensitive (age and gender) and perhaps a third feature that could also be considered sensitive (lives-near). One needs to decide based on domain knowledge what features and hence groups to consider. It is recommended to define multiple to be protected groups, by combining sensitive features.

Thirdly we have shown that some definitions are difficult to apply in practice due to practical challenges. Group fairness cannot be determined based on decision data alone. One needs access to the algorithm. This is for instance shown in Table 9 where it is unknown whether the expert method satisfies fairness through unawareness. We also discussed common metrics fair inference and individual fairness through similarity metrics and explained why these are not easy to apply.

Fourthly, it is remarkable that all of the fairness metrics do not use the information available about the value of each hire to the business or the loss of each candidate. All these metrics assume a candidate should be definitely hired or definitely not hired and do not account for the fact that some candidates are marginally suitable or unsuitable. This limited view makes it much hard to

optimize business value and societal value and makes some algorithms seem needlessly unfair. We would recommend the development of new metrics which take the differences in values into account, or that at least allow for decisions to be acceptable either way.

Fifthly, we would like to point out that fairness metrics are often introduced without clearly stating an acceptable tolerance. This makes it harder for auditors and implementers to reach an overall conclusion on fairness. Especially when you decide to do a thorough fairness review by considering multiple sensitive features, there is a risk of finding no algorithm acceptable.

This is shown in Table 11. The table shows a detailed analysis of multiple methods against multiple definitions considering only one subgroup. In all but one comparison in the table the metrics are not approximately equal, and one thus needs to call these algorithms unfair. For our Age>40 test scenario, with 78 over the total 225 candidates, the vast majority of fairness definitions are not satisfied by the classifiers, indicating that there is indeed some form of age-related bias in the data. The same occurs with Gender and 'lives near' attributes, where only few metrics are satisfied, and even more so for female candidates over 40 years old (Table 11).

6.2 Conclusion

In this work, we tackled the open problem of illustrating the various definitions of fairness metrics in machine learning. We have reviewed the relevant literature on fairness and provided a selection of relevant fairness definitions. The main contribution of this paper lies in illustrating these definitions by means of an artificial dataset of a realistic hiring scenario. The use of an artificial dataset is beneficial since it means that the dataset can be reused and shared without any restrictions or privacy risks. We conclude with a discussion on the contrasting results, the importance of the socio-technical context and domain knowledge for the application of fairness definitions.

We conclude that decision makers, algorithm auditors and supervisory authorities should apply relevant domain knowledge to assess how the various metrics align with business goals, legal obligations and company values and not blindly describe one metric and one to be protected group.

For instance, it might be more important to prioritize equality of odds (where truly qualified candidates have equal opportunity, and similarly, unqualified candidates face equal probabilities of being incorrectly selected across groups). Or on the other hand, group fairness (statistical parity) might be prioritized to improve demographic representation at the cost of the overall accuracy and business value. Every such situation has its own sensitive nuances that can lead to positive results but carry unwanted consequences, such as the case with positive discrimination policies.

6.3 Suggestions for further research

- There are numerous additional fairness definitions we have not covered in our examples, and disparate applicable scenarios where to make similar considerations to ours. For instance, among the reviewed literature, there remain metrics such as False negative error balance [Cho17, HPS16, KLRS17], (conditional) accuracy equality [BHJ⁺21], test-fairness [Cho17, HPS16], well-calibration [KMR16], balance for positive and negative class [KMR16], causal discrimination [GBM17], fairness through awareness [DHP⁺12], counterfactual fairness [KLRS17] and no unresolved and no proxy discrimination [KRCP⁺17]. It would be interesting to first investigate these metrics and then make a decision to determine the type of problems these metrics are suitable for.
- Other future work should be done to determine what people mean exactly when they want fairness in specific situations (e.g., candidate hires, fraud detection, but also price predictions or size estimations) using surveys, interviews or experiments. This could provide audit guidelines on which fairness definition to use.

- Most fairness research is focused on fairness in binary classification, assuming people are either completely suitable or completely unsuitable. Further research is needed into more nuanced metrics that would take into account that suitability for a function is a gradual concept.
- It would be helpful to have audit tools or guidelines on how to compute overall fairness over many different groups or using multiple metrics.

6.4 Conclusion

Acknowledgments

The authors thank Yiannis Kanellopoulos and Stefan Leijnen for their suggestions and feedback.

References

- [BDH⁺19] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, July 2019. Conference Name: IBM Journal of Research and Development.
- [BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.
- [BHJ⁺21] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021. <https://doi.org/10.1177/0049124118782533>.
- [Bin19] Reuben Binns. On the Apparent Conflict Between Individual and Group Fairness, December 2019. arXiv:1912.06883 [cs].
- [CCG⁺22] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, 2022. <https://doi.org/10.1038/s41598-022-07939-1>.
- [CDPF⁺17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017. <https://doi.org/10.1145/3097983.3098095>.
- [Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017. <https://doi.org/10.1089/big.2016.0047>.
- [CLM22] Hugo Cossette-Lefebvre and Jocelyn Maclure. AI’s fairness problem: understanding wrongful discrimination in the context of automated decision-making. *AI and Ethics*, pages 1–15, 2022.
- [Cod20] Code4Thought. Fairness, Accountability & Transparency (F.Acc.T) under GDPR, November 2020.
- [Cod25] Code4Thought. Code4thought - Trustworthy Solutions for AI, 2025.
- [Dee25] Deeploy. Deeploy explainable ML deployments - GitLab, February 2025.
- [DHP⁺12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. <https://doi.org/10.1145/2090236.2090255>.
- [DWWP24] Adam Dubowski, Hilde Weerts, Anouk Wolters, and Mykola Pechenzkiy. Subgroup Harm Assessor: Identifying Potential Fairness-Related Harms and Predictive Bias. In Albert Bifet, Povilas Daniušis, Jesse Davis, Tomas Krilavičius, Meelis Kull, Eirini Ntoutsi, Kai Puolamäki, and Indrė Žliobaitė, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track and Demo Track*, pages 413–417, Cham, 2024. Springer Nature Switzerland.
- [Eur16] Parliament European. General Data Protection Regulation (GDPR) Compliance Guidelines, 2016.
- [Fai22] Fairlearn. The Four Fifts Rule: Often Misapplied, 2022.

- [Fle21] Will Fleisher. What's Fair about Individual Fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 480–490, New York, NY, USA, July 2021. Association for Computing Machinery.
- [GBM17] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, pages 498–510, 2017. <https://doi.org/10.1145/3106237.3106277>.
- [Ger19] Aurelien Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 2nd edition, 2019.
- [Gui19] Ethics guidelines for trustworthy ai. 2019. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.
- [HM19] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 49–58, 2019. <https://doi.org/10.48550/arXiv.1811.10104>.
- [Hoo22] Maaik Hooghiemstra. Comparing different definitions of fairness in ai: A case study. 2022.
- [HPS16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. <https://doi.org/10.48550/arXiv.1610.02413>.
- [KLRS17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017. <https://doi.org/10.48550/arXiv.1703.06856>.
- [KMR16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016. <https://doi.org/10.48550/arXiv.1609.05807>.
- [KRCP⁺17] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017. <https://doi.org/10.48550/arXiv.1706.02744>.
- [Lee17] Thomas J. Leeper. Course materials for teaching R. https://fairlearn.org/main/user_guide/datasets/boston_housing_data.html, 2017. Accessed: 2025-02-27.
- [MWR23] Brent Mittelstadt, Sandra Wachter, and Chris Russell. The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. *arXiv preprint arXiv:2302.02404*, 2023. <https://doi.org/10.48550/arXiv.2302.02404>.
- [NS18] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. <https://doi.org/10.1609/aaai.v32i1.11553>.
- [OEC19] Oecd ai principles overview. 2019. <https://oecd.ai/en/ai-principles>.
- [PAKM22] Panagiotis Papagiannopoulos, Christos Aridas, Yiannis Kanellopoulos, and Christos Makris. Towards Discrimination-Free Classification via Fairness-Aware Bagging. In *Proceedings of the 25th Pan-Hellenic Conference on Informatics*, PCI '21, pages 184–189, New York, NY, USA, February 2022. Association for Computing Machinery.
- [PLL⁺23] Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*, 7(1):15, March 2023. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
- [PS19] Anya ER Prince and Daniel Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.*, 105:1257, 2019. <https://ssrn.com/abstract=3347959>.
- [SCDG17] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. 2017. <https://doi.org/10.1214/17-AOAS1058>.
- [SKH⁺19] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A Bias and Fairness Audit Toolkit, April 2019. arXiv:1811.05577 [cs].

- [SvO22] Piet Snel and Sieuwert van Otterloo. Practical bias correction in neural networks: a credit default prediction case study. *Computers and Society Research Journal*,(3), 2022. <https://doi.org/10.54822/BEWO3288>.
- [VR18] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018. <https://doi.org/10.1145/3194770.3194776>.
- [WB21] Jakub Wiśniewski and Przemysław Biecek. fairmodels: A flexible tool for bias detection, visualization, and mitigation. *arXiv preprint arXiv:2104.00507*, 2021. <https://doi.org/10.48550/arXiv.2104.00507>.
- [WDE⁺23] Hilde Weerts, Miroslav DudáĀk, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023.
- [WMC22] Elizabeth Anne Watkins, Michael McKenna, and Jiahao Chen. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness, February 2022. arXiv:2202.09519 [cs].
- [WMR21] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567, 2021. <https://doi.org/10.1016/j.clsr.2021.105567>.
- [Zli15] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1505.05723*, 2015. <https://doi.org/10.48550/arXiv.1511.00148>.
- [ZVRG17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017. <https://doi.org/10.48550/arXiv.1507.05259>.

Appendix 1. Appendix

Table 10 shows the dummy data used to generate the synthetic candidates in our dataset. For instance, the first row in the dataset (mainid 225000) is build by looking up the first rows of Tables 10a and 10b and fetching the respective Age, Gender, Speed, Speedtest, Strength and Lifttest. Once 15 such rows are fetched, a modulo 15 operation on the lookup operation restarts the counter to zero till 225 (15*15) dummy candidates are generated. As a final step, we add jitter (a randomized increment between 0 and 9) to the Age variable to provide a realistic feeling to the data.

Table 10. Dummy data used to generate the candidates dataset.

Id	Age	Gender	Speed	Speedtest	Id	Age	Gender	Strength	Lifttest
0	20	any	2	1	0	any	male	6	1
1	20	any	2	0.5	1	any	male	6	1
2	20	any	1	1	2	any	male	6	1
3	20	any	1	0.5	3	any	male	6	0.5
4	30	any	2	1	4	any	male	6	0.5
5	30	any	1	1	5	any	male	5	1
6	30	any	1	0.5	6	any	male	5	1
7	30	any	0	0.5	7	any	male	5	0.5
8	40	any	0	1	8	any	male	4	0.5
9	40	any	1	0.5	9	any	male	4	0
10	40	any	1	0.5	10	any	female	6	1
11	40	any	0	0.5	11	any	female	5	0.5
12	40	any	0	1	12	any	female	5	0.5
13	40	any	0	0.5	13	any	female	4	0.5
14	40	any	0	0	14	any	female	4	0

(a) Data for speed and speedtest variables (age-dependent) (b) Data for Strength and Lifttest variables (gender-dependent)

Figure 5 shows the distributions for relevant features such as Age (mean=35.8, std=8.5, min=20, max=49), Gender (female=75, male=150), testresult (mean=1.2, std=0.5, min=0.0, max=2.0) and Suitability (mean=1.9, std=1.1, min=1.0, max=4.0).

Table 11. Fairness metrics for protected attributes Gender and Age (p =female and Age>40, u =rest).

Method	Group	Group fairness	Pred. par. (PPV)	Pred. eq. (FPR)	Eqq. odds (TPR,FPR)	Treat. eq. (FN/FP)	Count
hired-by-expert	female>40	0.08	0.50	0.04	1.00, 0.04	0.00	24
hired-by-expert	rest	0.30	0.87	0.06	0.72, 0.06	2.62	201
Fair		X	X	✓	X	X	X
A1(testresult)	female>40	0.12	0.33	0.09	1.00, 0.09	0.00	24
A1(testresult)	rest	0.47	0.60	0.30	0.76, 0.30	0.47	201
Fair		X	X	X	X	X	X
A2(testresult,30under)	female>40	0.00	0.00	0.00	0.00, 0.00	inf	24
A2(testresult,30under)	rest	0.36	0.68	0.18	0.66, 0.18	1.09	201
Fair		X	X	X	X	X	X
A3(Age,Gender, test)	female>40	0.00	0.00	0.00	0.00, 0.00	inf	24
A3(Age,Gender, test)	rest	0.38	0.70	0.18	0.72, 0.18	0.91	201
Fair		X	X	X	X	X	X
A4(postive-dicr)	female>40	0.12	0.33	0.09	1.00, 0.09	0.00	24
A4(postive-dicr)	rest	0.29	0.55	0.20	0.43, 0.20	1.62	201
Fair		X	X	X	X	X	X

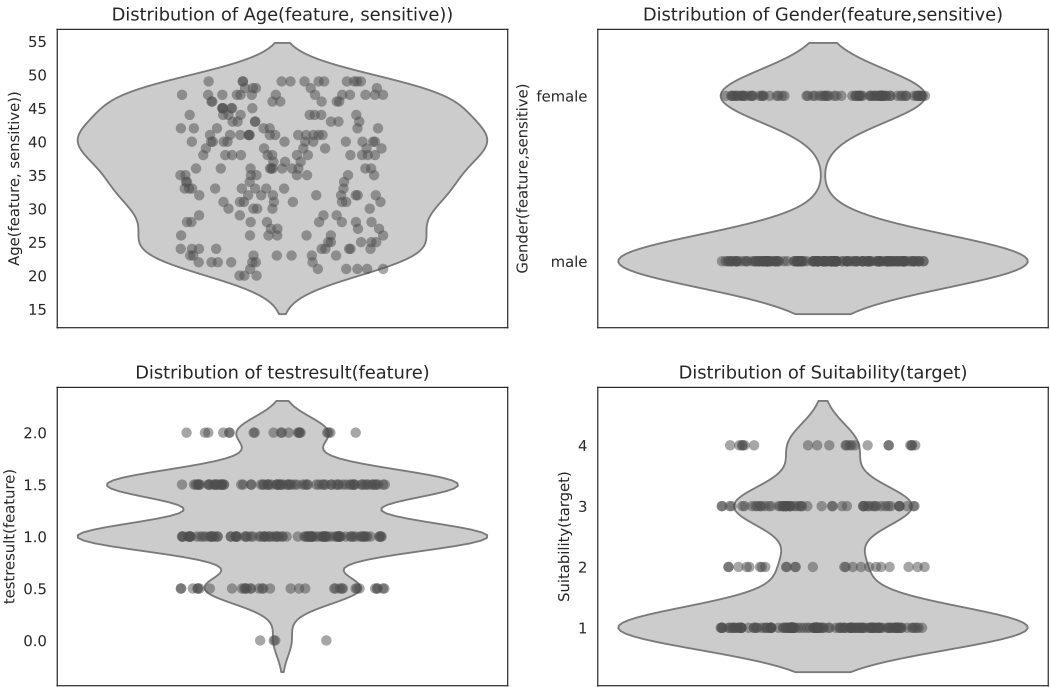


Figure 5. Distribution of Age, Gender, testresult and suitability attributes. There is notable prevalence of male applicants, slightly over 30 years old candidates. The testresult most often occurs between 1 and 1.5, and suitability is most often on 1.

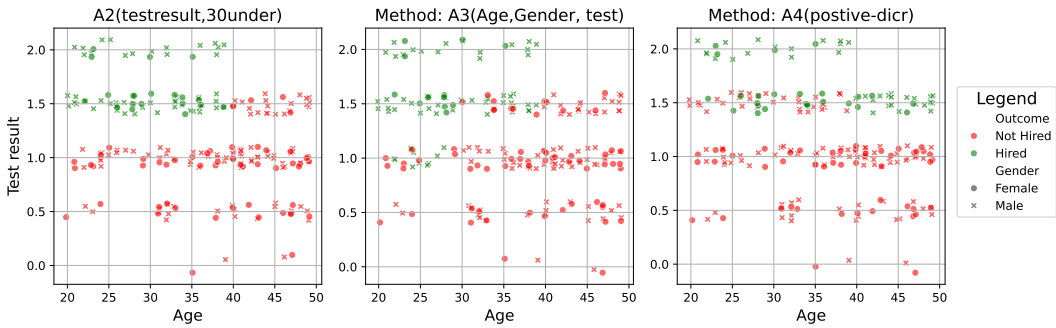


Figure 6. Scatterplot of hiring outcomes for A2, A3 and A4 against test score and age. The markers are coded by shape (female/male) and color (hired/not hired). The differences in hiring decision are evident by looking at the three methods discriminate by test result, age and gender.

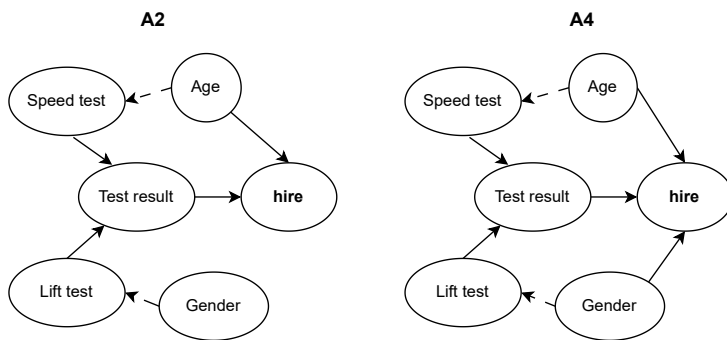


Figure 7. Causal graphs for A2 and A4 algorithms based on the dummy dataset generation. Nodes represent attributes ('hire' is the outcome variable) and edges relationships (dotted edge connects a proxy variable).