Fairness Definitions Explained

Sahil Verma
Indian Institute of Technology Kanpur, India
vsahil@iitk.ac.in

Julia Rubin
University of British Columbia, Canada
mjulia@ece.ubc.ca

ABSTRACT

Algorithm fairness has started to attract the attention of researchers in AI, Software Engineering and Law communities, with more than twenty different notions of fairness proposed in the last few years. Yet, there is no clear agreement on which definition to apply in each situation. Moreover, the detailed differences between multiple definitions are difficult to grasp. To address this issue, this paper collects the most prominent definitions of fairness for the algorithmic classification problem, explains the rationale behind these definitions, and demonstrates each of them on a single unifying case-study. Our analysis intuitively explains why the same case can be considered fair according to some definitions and unfair according to others.

ACM Reference Format:
Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *FairWare'18: IEEE/ACM International Workshop on Software Fairness, May 29, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3194770.3194776

1 INTRODUCTION

Recent years have brought extraordinary advances in the field of Artificial Intelligence (AI). AI now replaces humans at many critical decision points, such as who will get a loan [1] and who will get hired for a job [3]. One might think that these AI algorithms are objective and free from human biases, but that is not the case. For example, risk-assessment software employed in criminal justice exhibits race-related issues [4] and a travel fare aggregator steers Mac users to more expensive hotels [2].

The topic of algorithm fairness has begun to attract attention in the AI and Software Engineering research communities. In late 2016, the IEEE Standards Association published a 250-page draft document on issues such as the meaning of algorithmic transparency [6]; the final version of this document is expected to be adopted in 2019. The document covers methodologies to guide ethical research and design that uphold human values outlined in the U.N. Universal Declaration of Human Rights. Numerous definitions of fair treatment, e.g., [8, 10, 12, 14], were also proposed in academia. Yet, finding suitable definitions of fairness in an algorithmic context is a subject of much debate.

In this paper, we focus on the machine learning (ML) classification problem: identifying a category for a new observation given training data containing observations whose categories are known. We collect and clarify most prominent fairness definitions for classification used in the literature, illustrating them on a common, unifying example – the German Credit Dataset [18]. This dataset is commonly used in fairness literature. It contains information about 1000 loan applicants and includes 20 attributes describing each applicant, e.g., credit history, purpose of the loan, loan amount requested, marital status, gender, age, job, and housing status. It also contains an additional attribute that describes the classification outcome – whether an applicant has a good or a bad credit score.

When illustrating the definitions, we checked whether the classifier that uses this dataset exhibits gender-related bias. Our results were positive for some definitions and negative for others, which is consistent with earlier studies showing that some of the proposed definitions are mathematically incompatible [10, 11, 16]. The main contribution of this paper lies in an intuitive explanation and simple illustration of a large set of definitions we collected.

The remainder of the paper is structured as follows. Section 2 provides the necessary background and notations. Statistical, individual, and casual definitions of fairness are presented in Sections 3-5, respectively. We discuss lessons learned and outline ideas for future research in Section 6. Section 7 concludes the paper.

2 BACKGROUND

Considered Definitions. We reviewed publications in major conferences and journals on ML and fairness, such as NIPS, Big Data, AAAI, FATML, ICML, and KDD, in the last six years. We followed their references and also cross-validated our list with several reports that list known definitions of fairness [5, 7, 8, 21]. Most prominent definitions, together with the papers that introduce them and the number of citations for each paper on Google Scholar as of January 2018, is shown in the first four columns of Table 1.

Dataset. As our case study, we used German Credit Dataset [18]. Each record of this dataset has the following attributes:
1. Credit amount (numerical); 2. Credit duration (numerical); 3. Credit purpose (categorical); 4. Status of existing checking account (categorical); 5. Status of savings accounts and bonds (categorical); 6. Number of existing credits (numerical); 7. Credit history (categorical); 8. Installment plans (categorical); 9. Installment rate (numerical); 10. Property (categorical); 11. Residence (categorical); 12. Period of present residency (numerical); 13. Telephone (binary); 14. Employment (categorical); 15. Employment length (categorical); 16. Personal status and gender (categorical); 17. Age (numerical); 18. Foreign worker (binary); 19. Dependents (numerical); 20. Other debtors (categorical); 21. Credit score (binary).

For example, Alice is requesting a loan amount of 1567 DM for a duration of 12 months for the purpose of purchasing a television, with a positive checking account balance that is smaller than 200 DM, having less than 100 DM in savings account, and having one

FairWare'18, May 29, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5746-3/18/05…$15.00
https://doi.org/10.1145/3194770.3194776

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓
3.3.2	Well calibration	[16]	81	✓
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Table 1: Considered Definitions of Fairness

existing credit at this bank. She duly paid existing credits at the bank till now and has no other installment plan. She possesses a car and owns a house, has been living at the present residence for one year and has a registered telephone. She is a skilled employee, working in the present employment for past four years. She is a 22-year-old married female and is a German citizen. She has one dependent and no guarantors. The recorded outcome for Alice (attribute #21) is a good credit score.

We focus our illustration of fairness definitions on gender-related discrimination, i.e., whether male and female applicants are treated differently. The gender and the marital status of the applicants is specified in one attribute (attribute #16), which has five possible categorical values: single male, married male, divorced male, single female, married or divorced female. As in all the 1000 records of this dataset there is no case of a single female applicant, we focus our investigation on checking whether married and divorced males are treated differently than married and divorced females.

Notations. In the rest of the paper, we use the following notations:
– G: Protected or sensitive attribute for which non-discrimination should be established.
– X: All additional attributes describing the individual.
– Y: The actual classification result (here, good or bad credit score of an applicant as described in the dataset: attribute #21)
– S: Predicted probability for a certain classification c, $P(Y = c|G, X)$ (here, predicted probability of having a good or bad credit score).
– d: Predicted decision (category) for the individual (here, predicted credit score for an applicant – good or bad); d is usually derived from S, e.g., $d = 1$ when S is above a certain threshold.

For Alice in the example above, the probability of having a good credit score (S) as established by a classifier is 88%. Thus, the predicted score (d) is good, same as the actual credit score recorded in the database (Y). Next, we describe in detail each of the analyzed fairness definitions and its meaning in the context of the German Credit Dataset.

Attribute	Coefficient
Personal status and gender: single male	0.16
Personal status and gender: married male	-0.04
Personal status and gender: married/divorced female	-0.08
Personal status and gender: divorced male	-0.14

Table 2: Coefficients of gender-related features

3 STATISTICAL MEASURES

We start by describing statistical notions of fairness, which form the basis for other, more advanced definitions described later in the paper. For our discussion, we trained an off-the-shelf logistic regression classifier in Python. We applied the ten-fold cross-validation technique, using 90% of the data for training and the remaining 10% of the data for testing and illustrating each of the definitions. We used numerical and binary attributes directly as features in the classification and converted each categorical attribute to a set of binary features, arriving at 48 features in total.

We explicitly included the protected gender attribute in our training, as it appears to influence the predicted credit score: Table 2 lists coefficients learned by the classifier for all features derived from the personal status and gender attribute. The classifier appears to favor single males when deciding on the credit score and disadvantage divorced males. Female applicants seem to receive similar treatment as married male applicants. Looking at the coefficient, one might conclude that the classifier does not explicitly disadvantage female applicants. In the rest of the section, we explore whether married/divorced female applicants get unfair treatment comparing with married/divorced male applicants according to various definitions of fairness known from the literature. We focus on married/divorced applicants because the dataset does not contain instances of single females.

Statistical Metrics. Most statistical measures of fairness rely on the following metrics, which are best explained using a confusion matrix – a table that is often used in ML to describe the accuracy of a classification model [22]. Rows and columns of the matrix represent instances of the predicted and actual classes, respectively. For a binary classifier, both predicted and actual classes have two values: positive and negative (see Table 3). In our case study, positive and negative classes correspond to good and bad credit scores, respectively. Cells of the confusion matrix help explain the following definitions:

1. True positive (TP): a case when the predicted and actual outcomes are both in the positive class.

2. False positive (FP): a case predicted to be in the positive class when the actual outcome belongs to the negative class.

3. False negative (FN): a case predicted to be in the negative class when the actual outcome belongs to the positive class.

4. True negative (TN): a case when the predicted and actual outcomes are both in the negative class.

5. Positive predictive value (PPV): the fraction of positive cases correctly predicted to be in the positive class out of all predicted positive cases, $\frac{TP}{TP+FP}$. PPV is often referred to as precision, and represents the probability of a subject with a positive predictive value to truly belong to the positive class, $P(Y = 1|d = 1)$. In our example, it is the probability of an applicant with a good predicted credit score to actually have a good credit score.

6. False discovery rate (FDR): the fraction of negative cases incorrectly predicted to be in the positive class out of all predicted positive cases, $\frac{FP}{TP+FP}$. FDR represents the probability of false acceptance, $P(Y = 0|d = 1)$, e.g., the probability of an applicant with a good predicted credit score to actually have a bad credit score.

7. False omission rate (FOR): the fraction of positive cases incorrectly predicted to be in the negative class out of all predicted negative cases, $\frac{FN}{TN+FN}$. FOR represents the probability of a positive case to be incorrectly rejected, ($P(Y = 1|d = 0)$), e.g, the probability of an applicant with a bad predicted credit score to actually have a good score.

8. Negative predictive value (NPV): the fraction of negative cases correctly predicted to be in the negative class out of all predicted negative cases, $\frac{TN}{TN+FN}$. NPV represents the probability of a subject with a negative prediction to truly belong to the negative class, $P(Y = 0|d = 0)$, e.g., the probability of an applicant with a bad predicted credit score to actually have such score.

9. True positive rate (TPR): the fraction of positive cases correctly predicted to be in the positive class out of all actual positive cases, $\frac{TP}{TP+FN}$. TPR is often referred to as sensitivity or recall; it represents the probability of the truly positive subject to be identified as such, $P(d = 1|Y = 1)$. In our example, it is the probability of an applicant with a good credit score to be correctly assigned with such score.

10. False positive rate (FPR): the fraction of negative cases incorrectly predicted to be in the positive class out of all actual negative cases, $\frac{FP}{FP+TN}$. FPR represents the probability of false alarms – falsely accepting a negative case, $P(d = 1|Y = 0)$, e.g., the probability of an applicant with a actual bad credit score to be incorrectly assigned with a good credit score.

11. False negative rate (FNR): the fraction of positive cases incorrectly predicted to be in the negative class out of all actual positive cases, $\frac{FN}{TP+FN}$. FNR represents the probability of a negative result given an actually positive subject, $P(d = 0|Y = 1)$, e.g., the probability of an applicant with a good credit score to be incorrectly assigned with a bad credit score.

12. True negative rate (TNR): the fraction of negative cases correctly predicted to be in the negative class out of all actual negative cases, $\frac{TN}{FP+TN}$. TNR represents the probability of a subject from the negative class to be assigned to the negative class, $P(d = 0|Y = 0)$, e.g., the probability of an applicant with a bad credit score to be correctly assigned with such score.

Next, we list statistical definitions of fairness that are based on these metrics.

3.1 Definitions Based on Predicted Outcome

The definitions listed in this section focus on a predicted outcome d for various demographic distributions of subjects. They represent the simplest and most intuitive notion of fairness. Yet, they have several limitations addressed by definitions listed in later sections.

3.1.1. **Group fairness [12] (a.k.a. statistical parity [12], equal acceptance rate [24], benchmarking [9])**. A classifier satisfies this definition if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class. In our example, this would imply equal probability for male and female applicants to have good predicted credit score: $P(d = 1|G = m) = P(d = 1|G = f)$.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Table 3: Confusion matrix

The main idea behind this definition is that applicants should have an equivalent opportunity to obtain a good credit score, regardless of their gender. In our case study, the probability to have a good predicted credit score for married / divorced male and female applicants is 0.81 and 0.75, respectively. As it is more likely for a male applicant to have good predicted score, we deem our classifier to fail in satisfying this definition of fairness. We record our decision for each definition in the last column of Table 1.

3.1.2. **Conditional statistical parity [11]**. This definition extends the previous one by permitting a set of legitimate attributes to affect the outcome. The definition is satisfied if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class, controlling for a set of legitimate factors L. In our example, possible legitimate factors that affect an applicant creditworthiness could be the requested credit amount, applicant's credit history, employment, and age. Considering these factors, male and female applicants should have equal probability of having good credit score: $P(d = 1|L = l, G = m) = P(d = 1|L = l, G = f)$.

In our case study, when controlling for factors L listed above, the probability for married / divorced male and female applicants to have good predicted credit score is 0.46 and 0.49, respectively. Unlike in the previous definition, here a female applicant is slightly more likely to get a good predicted credit score. However, even though the calculated probabilities are not strictly equal, for practical purposes, we consider this difference minor, and hence deem the classifier to satisfy this definition.

3.2 Definitions Based on Predicted and Actual Outcomes

The definitions in this section not only consider the predicated outcome d for different demographic distributions of the classification subjects, but also compare it to the actual outcome Y recorded in the dataset.

3.2.1. **Predictive parity [10] (a.k.a. outcome test [9])**. A classifier satisfies this definition if both protected and unprotected groups have equal PPV – the probability of a subject with positive predictive value to truly belong to the positive class. In our example, this implies that, for both male and female applicants, the probability of an applicant with a good predicted credit score to actually have a good credit score should be the same:
$P(Y = 1|d = 1, G = m) = P(Y = 1|d = 1, G = f)$.

Mathematically, a classifier with equal PPVs will also have equal FDRs: $P(Y = 0|d = 1, G = m) = P(Y = 0|d = 1, G = f)$.

The main idea behind this definition is that the fraction of correct positive predictions should be the same for both genders. In our

case study, PPV for married / divorced male and female applicants is 0.73 and 0.74, respectively. Inversely, FDR for male and female applicants is 0.27 and 0.26, respectively. The values are not strictly equal, but, again, we consider this difference minor, and hence deem the classifier to satisfy this definition.

3.2.2. **False positive error rate balance [10] (a.k.a. predictive equality [11])**. A classifier satisfies this definition if both protected and unprotected groups have equal FPR – the probability of a subject in the negative class to have a positive predictive value. In our example, this implies that the probability of an applicant with an actual bad credit score to be incorrectly assigned a good predicted credit score should be the same for both male and female applicants: $P(d = 1|Y = 0, G = m) = P(d = 1|Y = 0, G = f)$.

Mathematically, a classifier with equal FPRs will also have equal TNRs: $P(d = 0|Y = 0, G = m) = P(d = 0|Y = 0, G = f)$.

The main idea behind this definition is that a classifier should give similar results for applicants of both genders with actual negative credit scores. In our case study, FPR for married / divorced male and female applicants is 0.70 and 0.55, respectively. Inversely, TNR is 0.30 and 0.45. This means that the classifier is more likely to assign a good credit score to males who have an actual bad credit score; females do not have such an advantage and the classifier is more likely to predict a bad credit score for females who actually have a bad credit score. We thus deem our classifier to fail in satisfying this definition of fairness.

3.2.3. **False negative error rate balance [10] (a.k.a. equal opportunity [14, 17])**. A classifier satisfies this definition if both protected and unprotected groups have equal FNR – the probability of a subject in a positive class to have a negative predictive value. In our example, this implies that the probability of an applicant with an actual good credit score to be incorrectly assigned a bad predicted credit score should be the same for both male and female applicants: $P(d = 0|Y = 1, G = m) = P(d = 0|Y = 1, G = f)$.

Mathematically, a classifier with equal FNRs will also have equal TPR: $P(d = 1|Y = 1, G = m) = P(d = 1|Y = 1, G = f)$.

The main idea behind this definition is that classifier should give similar results for applicants of both genders with actual positive credit scores. In our case study, the FPRs for married / divorced male and female applicants are the same – 0.14. Inversely, TPR is 0.86. Like in the case of predictive parity (3.2.1), this means that the classifier will apply equivalent treatment to male and female applicants with actual good credit score. We thus deem our classifier to satisfying this definition of fairness.

If the prevalence of a good credit score is the same for male and female subjects in the entire population, this definition becomes equivalent to the *group fairness* definition (3.1.1) which requires equal probability for male and female applicants to have a good predicted credit score. Yet, in general, the definitions are not equivalent [8, 10, 16]. In our example, male applicants in the studied population are more likely to have a good actual credit score. Thus, the classifier is also more likely to assign a good predicted credit score to male applicants. For that reason, our classifier satisfies the *equal opportunity* but does not satisfy the *group fairness* definitions.

3.2.4. **Equalized odds [14] (a.k.a. conditional procedure accuracy equality [8] and disparate mistreatment [23])**. This definition combines the previous two: a classifier satisfies the definition if protected and unprotected groups have equal TPR and equal FPR. Mathematically, it is equivalent to the conjunction of conditions for false positive error rate balance and false negative error rate balance definitions given above. In our example, this implies that the probability of an applicant with an actual good credit score to be correctly assigned a good predicted credit score and the probability of an applicant with an actual bad credit score to be incorrectly assigned a good predicted credit score should both be same for male and female applicants: $P(d = 1|Y = i, G = m) = P(d = 1|Y = i, G = f), i \in 0, 1$.

The main idea behind this definition is that applicants with a good actual credit scope and applicants with a bad actual credit score should have a similar classification, regardless of their gender. In our case study, FPR for married / divorced male and female applicants is 0.70 and 0.55, respectively and TPR is 0.86 for both males and females. This means that the classifier is more likely to assign a good credit score to males who have an actual bad credit score, compared to females. Hence the overall conjunction does not hold and we deem our classifier to fail in satisfying this definition.

If male and female applicants have different probabilities to be in the actual positive class $P(Y = 1|G = m) \neq P(Y = 1|G = f)$, a classifier that satisfies *predictive parity* (3.2.1) cannot satisfy this definition [10]. Our observations are consistent with that theoretical result.

3.2.5. **Conditional use accuracy equality [8]**. Similar to the previous definition, this definition conjuncts two conditions: equal PPV and NPV – 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: $(P(Y = 1|d = 1, G = m) = P(Y = 1|d = 1, G = f)) \wedge (P(Y = 0|d = 0, G = m) = P(Y = 0|d = 0, G = f))$.

Intuitively, this definition implies equivalent accuracy for male and female applicants from both positive and negative predicted classes. In our example, the definition implies that for both male and female applicants, the probability of an applicant with a good predicted credit score to actually have a good credit score and the probability of an applicant with a bad predicted credit score to actually have a bad credit score should be the same. The calculated for male and female applicants is 0.73 and 0.74, respectively. NPVs for male and female applicants is 0.49 and 0.63 respectively. It is more likely for a male than female applicant with a bad predicted score to actually have a good credit score. We thus deem the classifier to fail in satisfying this definition of fairness.

3.2.6. **Overall accuracy equality [8]**. A classifier satisfies this definition if both protected and unprotected groups have equal prediction accuracy – the probability of a subject from either positive or negative class to be assigned to its respective class. The definition assumes that true negatives are as desirable as true positives. In our example, this implies that the probability of an applicant with an actual good credit score to be correctly assigned a good predicted credit score and an applicant with an actual bad credit score to be correctly assigned a bad predicted credit score is the same for both male and female applicants: $P(d = Y, G = m) = P(d = Y, G = f)$.

In our case study, the overall accuracy rate is 0.68 and 0.71 for male and female applicants, respectively. While these values are not strictly equal, for practical purposes we consider this difference minor, and hence deem the classifier to satisfy this definition. This means that the classifier has equal prediction accuracy for both

s	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$P(Y = 1\|S = s, G = m)$	1.0	1.0	0.3	0.3	0.4	0.6	0.6	0.7	0.8	0.8	1.0
$P(Y = 1\|S = s, G = f)$	0.5	0.3	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Table 4: Calibration scores for different values of s

genders when particular classes of subjects (e.g., subjects with positive predicted class) are not considered separately.

3.2.7. **Treatment equality [8]**. This definition looks at the ratio of errors that the classifier makes rather than at its accuracy. A classifier satisfies this definition if both protected and unprotected groups have an equal ratio of false negatives and false positives. In our example, this implies that the ratio of FP to FN is same for male and female applicants: $\frac{FN}{FP} m = \frac{FN}{FP} f$. This calculated ratios are 0.56 and 0.62 for male and female applicants, respectively, i.e., a smaller number of male candidates are incorrectly assigned to the negative class (FN) and / or larger number of male candidates are incorrectly assigned to the positive class (FP). We thus deem our classifier to fail in satisfying this definition of fairness.

3.3 Definitions Based on Predicted Probabilities and Actual Outcome

The definitions in this section consider the actual outcome Y and the predicted probability score S.

3.3.1. **Test-fairness [10] (a.k.a. calibration [10], matching conditional frequencies [14])**. A classifier satisfies this definition if for any predicted probability score S, subjects in both protected and unprotected groups have equal probability to truly belong to the positive class. This definition is similar to *predictive parity* (3.2.1), except that it considers the fraction of correct positive predictions for any value of S.

In our example, this implies that for any given predicted probability score s in $[0, 1]$, the probability of having actually a good credit score should be equal for both male and female applicants: $P(Y = 1|S = s, G = m) = P(Y = 1|S = s, G = f)$.

In our case study, we calculated the predicted score S for each applicant in the test set, and binned the results in 11 bins, from 0.0 to 1.0. Table 4 shows the scores for male and female applicants in each bin. The scores are quite different for lower values of S and become closer for values greater than 0.5. Thus, our classifier satisfies the definition for high predicted probability scores but does not satisfy it for low scores. This is consistent with previous results showing that it is more likely for a male applicant with a bad predicted credit score (low values of S) to actually have a good score (definition 3.2.5), but applicants with a good predicted credit score (high values of S) have an equivalent chance to indeed have a good credit score, regardless of their gender (definition 3.2.1).

3.3.2. **Well-calibration [16]**. This definition extends the previous one stating that, for any predicted probability score S, subjects in both protected and unprotected groups should not only have an equal probability to truly belong to the positive class, but this probability should be equal to S. That is, if the predicted probability score is s, the probability of both male and female applicants to truly belong to the positive class should be s. $P(Y = 1|S = s, G = m) = P(Y = 1|S = s, G = f) = s$.

The intuition behind this definition is that if a classifier states that a set of applicants have a certain probability s of having a good credit score then approximately s percent of these applicants should indeed have a good credit score. In our case study, scores for male and female applicants calculated for each value of s are binned and shown in Table 4. Our classifier is well-calibrated only for $s \geq 0.6$. We thus deem the classifier to partially satisfy this fairness definition.

3.3.3. **Balance for positive class [16]**. A classifier satisfies this definition if subjects constituting positive class from both protected and unprotected groups have equal average predicted probability score S. Violation of this balance means that one group of applicants with good credit score would consistently receive higher probability score than applicants with a good credit score from the other group.

In our example, this implies that the expected value of probability assigned by the classifier to male and female applicant with good actual credit score should be same: $E(S|Y = 1, G = m) = E(S|Y = 1, G = f)$. The calculated expected value of predicted probability score is 0.72 for both males and females and we thus deem the model to satisfy this notion of fairness. This result further supports and is consistent with the result for *equal opportunity* (3.2.3), which states that the classifier will apply equivalent treatment to male and female applicants with actual good credit score (TPR of 0.86).

3.3.4. **Balance for negative class [16]**. In a flipped version of the previous definition, this definition states that subjects constituting negative class from both protected and unprotected groups should also have equal average predicted probability score S. That is, the expected value of probability assigned by the classifier to male and female applicant with bad actual credit score should be same: $E(S|Y = 0, G = m) = E(S|Y = 0, G = f)$.

In our case study, the expected value of having bad predicted credit score is 0.61 and 0.52 for males and females, respectively. This means that, on average, male candidates who actually have bad credit score receive higher predicted probability scores than female candidates. We thus deem our classifier to fail in satisfying this definition of fairness. This result further supports and is consistent with the result for *predictive equality* (3.2.2), which states that the classifier is more likely to assign a good credit score to males who have an actual bad credit score (TNR of 0.30 and 0.45 for males and females, respectively).

4 SIMILARITY-BASED MEASURES

Statistical definitions largely ignore all attributes of the classified subject except the sensitive attribute G. Such treatment might hide unfairness: suppose the same fraction of male and female applicants are assigned a positive score. Yet, male applicants in this set are chosen at random, while female applicants are only those that have the most savings. Then, statistical parity will deem the classifier fair, despite a discrepancy in how the applications are processed based on gender [13]. The following definitions attempt to address such issues by not marginalizing over insensitive attributes X of the classified subject.

4.1. **Causal discrimination [13]**. A classifier satisfies this definition if it produces the same classification for any two subjects with the exact same attributes X. In our example, this implies that a male and female applicants who otherwise have the same attributes X will either both be assigned a good credit score or both assigned a bad credit score: $(X_f = X_m \wedge G_f \; != G_m) \rightarrow d_f = d_m$.

To test this definition for our case study, for each applicant in our testing set, we generated an identical individual of the opposite gender and compared the predicted classification for these two applicants. We found that for 8.8% married / divorced male and female applicants, the output classification was not same. We thus deem our classifier to fail in satisfying this definition.

4.2. **Fairness through unawareness [17]**. A classifier satisfies this definition if no sensitive attributes are explicitly used in the decision-making process. In our example, this implies that gender-related features are not used for training the classifier, so decisions cannot rely on these features. This also means that the classification outcome should be the same for applicants i and j who have the same attributes X: $X_i = X_j \rightarrow d_i = d_j$.

To test this definition for our case study, we trained the logistic regression model without using any features derived from the gender attribute. Then, for each applicant in the testing set, we generated an identical individual of the opposite gender and compared the predicted classification for these two applicants. Our results show that the classification for all "identical" individuals that only differ in gender was identical. We thus deem the classifier to satisfy this definition. This result also indicates that no other feature of the dataset is used as a proxy for gender; otherwise, the classifier would have shown similar results as in case of causal discrimination.

4.3. **Fairness through awareness [12]**. This definition is a more elaborated and generic version of the previous two: here, fairness is captured by the principle that similar individuals should have similar classification. The similarity of individuals is defined via a distance metric; for fairness to hold, the distance between the distributions of outputs for individuals should be at most the distance between the individuals. Formally, for a set of applicants V, a distance metric between applicants $k : V \times V \rightarrow R$, a mapping from a set of applicants to probability distributions over outcomes $M : V \rightarrow \delta A$, and a distance D metric between distribution of outputs, fairness is achieved iff $D(M(x), M(y)) \leq k(x, y)$.

For example, a possible distance metric k could define the distance between two applicants i and j to be 0 if the attributes in X (all attributes other than gender) are identical and 1 if some attributes in X are different. D could be defined as 0 if the classifier resulted in the same prediction and 1 otherwise. This basically reduces the problem to the definition of causal discrimination (4.1), and the same result holds: for 8.8% of the applicants the fairness constraint is violated.

As another example, the distance metric between two individuals could be defined as the normalized difference of their ages: the age difference divided by the maximum difference in the dataset (56 in our case). The distance between outcomes could be defined as the statistical difference between the outcome probabilities for two applicants: $D(i, j) = S(i) - S(j)$.

To test this definition, for each applicant in the testing set, we generated five additional individuals, with ages different by 5, 10, 15, 20 and 25 years, and identical otherwise. Our results in Table 5 show that the distance between outcomes (column 3) grew much faster than the distance between ages (column 3). Thus, the percentage of applicants who did not satisfy this definition (column 4) increased. That is, for a smaller age difference, the classifier satisfied this fairness definition, but that was not the case for an age difference of more than 10 years. This result also shows that a distance metric is of fundamental importance when applying this definition and should be chosen with care.

Age difference	k	Avg. D	% violating cases
5	0.09	0.02	0.0
10	0.18	0.05	0.5
15	0.27	0.10	1.8
20	0.36	0.2	4.5
25	0.45	0.3	6.7

Table 5: Fairness through awareness with age-based distance

5 CAUSAL REASONING

Definitions based on causal reasoning assume a given causal graph: a directed, acyclic graphs with nodes representing attributes of an applicant and edges representing relationships between the attributes. Causal graphs are used for building fair classifiers and other ML algorithms [15, 17, 19, 20]. Specifically, the relations between attributes and their influence on outcome is captured by a set of structural equations which are further used to provides methods to estimate effects of sensitive attributes and build algorithms that ensure a tolerable level of discrimination due to these attributes.

While it is impossible to test an existing classifier against causal definitions of fairness, we demonstrate them on a simple causal graph we built for our dataset for illustration purposes. Our graph (see Figure 1) consists of the protected attribute G, the credit amount, employment length, and credit history attributes, and the predicted outcome d.

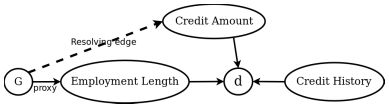

Figure 1: Causal graph example

In causal graphs, a *proxy* attribute is an attribute whose value can be used to derive a value of another attribute. In our example, we assume that employment length acts as a proxy attribute for G: one can derive the applicants' gender from the length of their employment.

A *resolving* attribute is an attribute in the causal graph that is influenced by the protected attribute in a non-discriminatory manner. In our example, the effect of G on the credit amount is non-discriminatory, which means that the differences in credit amount for different values of G are not considered as discrimination. Hence, the credit amount acts as a resolving attribute for G in this graph.

5.1. **Counterfactual fairness [17]**. A causal graph is counterfactually fair if the predicted outcome d in the graph does not depend on a descendant of the protected attribute G. For the example in Figure 1, d is a dependent on credit history, credit amount, and employment length. Employment length is a direct descendant of G, hence, the given causal model is not counterfactually fair.

5.2. **No unresolved discrimination [15]**. A causal graph has no unresolved discrimination if there exists no path from the protected attribute G to the predicted outcome d, except via a resolving variable. In our example, the path from G to d via credit amount is non-discriminatory as the credit amount is a resolving attribute; the path via employment length is discriminatory. Hence, this graph exhibits unresolved discrimination.

5.3. **No proxy discrimination [15]**. A causal graph is free of proxy discrimination if there exists no path from the protected attribute G to the predicted outcome d that is blocked by a proxy variable. For the example in Figure 1, there is an indirect path from G to d via proxy attribute employment length. Thus, this graph exhibits proxy discrimination.

5.4. **Fair inference [19]**. This definition classifies paths in a causal graph as legitimate or illegitimate. For example, it might make sense to consider the employment length for making credit-related decision. Even though the employment length acts as a proxy for G, that path would be considered as legitimate. A causal graph satisfies the notion of fair inference if there are no illegitimate paths from G to d, which is not the case in our example as there exist another illegitimate path, via credit amount.

6 DISCUSSION AND LESSONS LEARNED

We observed that a logistic regression classifier trained on the German Credit Dataset is more likely to assign a good credit score to male applicants in general (3.1.1) and male applicants who have an actual bad credit score in particular (3.2.2 and 3.2.4). Females do not have such an advantage and the classifier is more likely to predict a bad credit score for females who have an actual bad credit score (3.2.2 and 3.2.4). Yet, the classifier applies equivalent treatment to male and female applicants with actual good credit score (3.2.3). It is also accurate in the sense that the probability of an applicant with an actual good (bad) credit score to be correctly assigned a good (bad) predicted credit score is the same for both male and female applicants (3.2.6). At the same time, it is more likely for a male applicant with a bad predicted score to have an actual good credit score (3.2.5), so the classifier disadvantages some "good" male applicants. Our results also show that the outcome for otherwise "identical" male and female applicants was not the same (4.1), but this problem disappears when the classifier was trained without considering gender-related attributes (4.2).

So, is the classifier fair? Clearly, the answer to this question depends on the notion of fairness one wants to adopt. We believe more work is needed to clarify which definitions are appropriate to each particular situation. We intend to make a step in this direction by systematically analyzing existing reports on software discrimination, identifying the notion of fairness employed in each case, and classifying the results.

A statistical notion of fairness as described in Section 3 is easy to measure. However, it was shown that statistical definitions are insufficient [8, 10, 12, 16]. Moreover, most valuable statistical metrics assume availability of actual, verified outcomes. While such outcomes are available for the training data, it is unclear whether the real classified data always conforms to the same distribution.

More advanced definitions discussed in Section 4 and 5 require expert input and opinion, e.g., to establish a distance metric between individuals. Not only are these definitions more difficult to measure, they can still be biased given implicit biases of the expert.

Finally, testing several definitions, such as *fairness through awareness*, relies on availability of "similar" individuals. Generating all possible data for testing such definition is clearly impractical as the search space could be very large (e.g., the global population). More work to narrow down the search space without impeding the accuracy of the analysis is needed.

7 CONCLUSIONS

In this paper, we collected most prominent definitions of fairness for the algorithmic classification problem. We explained and demonstrated each definition of a single unifying example of an off-the-shelf logistic regression classifier trained on the German Credit Dataset. The main contribution of this paper lies in the intuitive explanation of each definition and identification of relationships between the definitions. We discussed lessons learned from our experiments and proposed directions for possible future work.

REFERENCES

[1] 2011. The Algorithm That Beats Your Bank Manager. https://www.forbes.com/sites/parmyolson/2011/03/15/the-algorithm-that-beats-your-bank-manager/. (March 2011). Online; accessed February 2018.

[2] 2012. On Orbitz, Mac Users Steered to Pricier Hotels. https://www.wsj.com/articles/SB10001424052702304458604577488822667325882. (August 2012). Online; accessed February 2018.

[3] 2015. Can an Algorithm Hire Better Than a Human? https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html. (June 2015). Online; accessed February 2018.

[4] 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. (May 2016). Online; accessed February 2018.

[5] 2017. CS 294: Fairness in Machine Learning. https://fairmlclass.github.io. (August 2017). Online; accessed February 2018.

[6] 2017. Ethically Aligned Design: A Vision for Prioritizing Human Well-being With Artificial Intelligence and Autonomous Systems. http://standards.ieee.org/develop/indconn/ec/auto_sys_form.html. (December 2017). Online; accessed February 2018.

[7] 2017. Fairness. https://speak-statistics-to-power.github.io/fairness/. (December 2017). Online; accessed January 2018.

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth, Richard Berka, Hoda Heidaric, Shahin Jabbaric, Michael Kearnsc, and Aaron Rothc. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art.

[9] Simoiu Camelia, Corbett-Davies Sam, and Goel Sharad. 2017. The Problem of Infra-marginality in Outcome Tests for Discrimination. *Ann. Appl. Stat. Vol. 11, No. 3* (2017).

[10] Alexandra Chouldechova. 2016. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* (2016).

[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proc. KDD'17.*

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference.*

[13] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness Testing: Testing Software for Discrimination. In *Proc. of ESEC/FSE'17.*

[14] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems.*

[15] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. 2017. Avoiding Discrimination Through Causal Reasoning. In *Advances in Neural Information Processing Systems.*

[16] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS.*

[17] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems.*

[18] M. Lichman. 2013. UCI Machine Learning Repository. (2013). http://archive.ics.uci.edu/ml

[19] R. Nabi and I. Shpitser. 2018. Fair Inference On Outcomes. In *AAAI.*

[20] Judea Pearl. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York, NY, USA.

[21] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems.*

[22] Foster Provost and Ron Kohavi. 1998. On Applied Research in Machine Learning. In *Machine learning.*

[23] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proc. of WWW'17.*

[24] Indre Zliobaite. 2015. On the Relation between Accuracy and Fairness in Binary Classification. *CoRR* abs/1505.05723 (2015).