Comparing Different Definitions of Fairness in AI: A Case Study

M.E. Hooghiemstra

Vrije Universiteit Amsterdam

Abstract. The prevalence of the help of AI in decision making processes is increasingly visible in society. The focus is also shifting towards the fairness of these collaborative AIs. In recent years, many relevant AI tooling have demonstrated to make discriminating judgement against specific social groups. For this research, a literature research on various fairness definitions is conducted. We also propose a recruitment model which decide whether applicants should be recruited or not. This prediction model is made such that the fairness definitions can be investigated on this recruitment use-case. Subsequently, this model is enhanced by means of post-processing. This will lead to a ranking of the fairness definitions regarding this model. Furthermore, a field survey is conducted within three society groups involved in a AI collaborative recruitment process; recruiters, potential applicants and AI experts. The survey concludes upon which definition are found to be important for a fair recruitment model. It also gives interesting insides on how testing of AI models is pledged to be done.

Keywords: Artificial Intelligence, Machine Learning, Fairness, Bias, Recruitment

1 Introduction

Nowadays, it is nearly impossible to think of our lives without the aid or influence of artificial intelligence (AI). AI is used when for example unlocking your phone with Face-ID, asking Siri for today's weather or when you are figuring out what to watch on your preferred streaming apps. But there are also many more areas in which AI is impacting our society, and not always in a positive way. One such example is in the area of collaborative intelligence, where AI models are used in combination with human input. This often involves classification or prediction, where the AI classifies or predicts unseen data and a human makes the final decision based on the input from the AI model. This sounds very promising in theory, but the practicality has some downsides that are not obvious from first glance. Due to its input, an AI algorithm is likely to adopt discrimination in its model. Examples of unfair discrimination in models are Amazon's hiring tool disadvantaging women [1], facial recognition software performing poor on black women [2] or even the fairly recent Dutch "toeslagenaffaire" [3]. These are examples of discrimination that had a huge impact on society, which makes it a highly interesting topic to research.

1.1 Motivation

As stated before, there are already some examples of AI models behaving poorly when faced with the practical sides of reality. Many recent news articles imply that AI models are simply still not “there” yet [4] [5] [6]. But what is “there” exactly? For many, “there” means that an AI model, among others, should not cause harm to historically disadvantaged groups, such as females, disabled people or people of colour. This involves banning racism, sexism and classicism from AI models such that trust in AI systems is (re)gained.

It is often forgotten that AI models do not come up with discrimination by themselves. They are simply applying their input to their algorithm and providing us with the output. As the discrimination of predictive AI models is often only noticed after the model has provided its outputs, it can be very difficult to see where the unfairness in the model stems from. There are many real-world examples of models that discriminate based on race, gender, age and social class. They show that this discrimination in the model can lead to unfair treatment of individuals, which can have extreme harmful impact on these individuals' lives. An example that was already mentioned is the Dutch “toeslagenaffaire” [3]. In this scandal, thousands of people had to pay back the Dutch taxing authority because the algorithm used incorrectly marked them as fraudulent. The used AI system discriminated based on ethnicity. Another example is ad advertisement from Instant Checkmate and Google [7]. It shows that online ads are discriminating on black-sounding names and white-sounding names.

As mentioned before, facial ID also uses AI to unlock a gadget. It requires many different faces to train the algorithm, such that it can distinguish between various features and determine the importance. However, when feeding the model only white faces as input, the facial recognition will perform significantly worse for women of colour, which negatively impacts their use of the feature [2] [8].

An example that also shows discrimination is a recruitment model of American e-commerce specialist Amazon that was implemented within the organization a few years back [1]. The model was trying to predict the best employee candidates for its company, but was found to be extremely discriminatory against female applicants. This eventually led to the termination of the complete predictive hiring model. This example shows that there was little insight and investigation in the ethical side of AI algorithms within an organization.

Over the past few decades, there is growing awareness of fairness in the machine learning domain. Above's real-world examples show that this awareness is needed to make AI models fairer, and to understand what is causing the unfairness in the model. To date, the interpretation of the assessment of prediction models has been the focal point of researchers [9]. While fairness is a legal obligation, this is not always achieved as seen in the examples mentioned above.

As already mentioned a lot in this introduction, fairness is important in AI models. But what exactly is fairness? A lot of different definitions are mentioned in the literature [10] [11]. Whenever bias is addressed fairness is quickly disregarded. For example, Snel demonstrates practical methods for reducing bias without loss of accuracy, but does not address fairness questions beyond this bias [12].

Furthermore, the OECD Policy Observation gives guidelines on how to AI in a just manner [13]. However, the literature does not state many use-cases of AI fairness related to recruiting.

1.2 Problem definition

A predictive model should be fair with regard to its outcome and the factors it can influence. AI models try to model real-life situations, correctly predict future possible cases and make well-grounded choices based on its evaluations. An AI model bases its predictions on the input data. Therefore it is crucial that sensitive characteristics, such as gender, sex and race, are carefully handled with respect to discrimination to achieve a fair model. As there is still very much debate on (degrees of) fairness, the goal of this research is to further examine the definition of a fair model from different kind of viewpoints namely from a mathematical, societal and end-user point of view. In this thesis, the focus will be on the mathematical and end-user points of view.

There is a lack of practical use-cases that give a guideline and understanding of fairness in AIs. This is evidenced by the lack of available datasets related to fairness on online coding websites such as Kaggle or Github.

There are many different fairness definitions as mentioned by, among others, Verma and Rubin [10] and Huang et al. [14]. For example, the statistical parity defintion state that equal fractions of each group should be treated as belonging to the positive class. Whereas, for example, the equal treatment definition state that the protected and unprotected group should have the same error rates. This results in ranging definitions of fairness, which makes defining fairness even more difficult. The problem is therefore the multiple definitions of fairness, and which one is the “most fair”.

1.3 Research question

This thesis proposes to explore the definitions of fairness, such that prediction models can move towards more fair models. To explore fairness, this research attempts to simulate real-world situations by using a realistic data set; a dataset that models a real-world situation and could be applied in a real-world application. To this end, the research attempts to answer the following research questions:

- **RQ:** How can fairness be measured and improved on a realistic data set?

In order to answer this main research question, we will answer the following sub research questions with respect to a recruitment use-case:

- **Sub-RQ 1:** What are the possible definitions of fairness with respect to recruitment processes?
- **Sub-RQ 2:** How to measure fairness on a trained model?

- **Sub-RQ 3:** What are techniques for reducing unfairness and bias relating to the fairness definitions?
- **Sub-RQ 4:** How will applying these techniques affect the recruitment predictive model?
- **Sub-RQ 5:** To what extend does society's interpretation of fairness up to par with the fairness outcomes of the reduced unfairness and bias recruitment model?

1.4 Scientific and practical contribution

Theoretical contribution. This research aims to extend the existing literature with further insights into the definition of fairness. Additionally, it provides insights in how fairness can best be measured in practical, real-world models. Consequently, it gives an indication for the definitions of fairness that are best used in future AI models.

Practical contribution. Building on the theoretical contribution, this research aims to give a better understanding of where fairness challenges within AI models lie. Additionally, it aims to provide a guideline on how to measure and improve fairness within a ML model, through a use-case. This utilization of a use-case will help organizations to enhance fairness within the model and company, and subsequently get a better understanding of their models. The code accompanying this project will be also be uploaded to give practical guidance to organizations.

2 Related literature

Societal movements and global trends, such as the BlackLivesMatter movement [15], dictate the need for equality in society. These movements and trends translate over to one of the biggest emerging technology trends in the world: equality in AI. Therefore, as mentioned in the introduction, literature on fairness in predictive modelling is an emerging field in machine learning. The meaning of fairness and bias, the kinds of fairness and the consequences in AI applications will be discussed in this literature section.

2.1 Fairness in society

In pursuit of fairness, laws prohibiting discrimination based on protected characteristics have been enacted for a number of high-stakes sectors, including credit, housing, education, and employment. According to Article 1 of the Dutch Constitution, everyone in the Netherlands must be treated equally [16]. Discrimination on the grounds of religion, belief, political opinion, race, sex or any other ground is prohibited.

Protected variables. Current legislation states a list of personal characteristics upon which it is obliged to never discriminate [16]. These are listed below.

- Religion/philosophy,
- Political affiliation,
- Race/origin,
- Gender: male, female, transgender (transsexuals, transvestites, intersex people),
- Pregnancy,
- Nationality,
- Hetero- or homosexual orientation (bisexual orientation), and
- Marital status: married or unmarried, with or without registered partnership.

Following from this, variables that can be discriminated upon are variables such as age, income and work experience.

Types of illegal discrimination. Current legislation recognizes two types of illegal discrimination, namely disparate treatment and disparate impact. Disparate treatment is the intentional discrimination of groups of people. It is sometimes referred to as purposeful prejudice. The treatment disproportionately harms people with specific sensitive attributes (e.g. gender, race). For example, when men are payed a higher salary than women for performing the same job. Whereas, disparate impact is unintentional discrimination. This is sometimes referred to as inadvertent discrimination. It arises when seemingly neutral policies, practices or other systems have a disproportionate influence on a protected group. For example, hiring more men than women as construction workers due to their physical strength.

2.2 AI applications and ethical questions

As already touched upon briefly, AI models use machine learning to classify, cluster or recognize patterns and predict in a wide range of domains [17]. For example, artificial neural networks are used for visual perception, language processing and other classification projects.

Many of these models are made to take over human tasks, therefore bias in these models is frequently discussed by society and academia. Bias is referred to as "computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favour of others" [18]. Bias can lead to negative or positive discrimination, where positive discrimination results in favourable treatment of one group and negative gives less favourable treatment of one group. The goal of fairness, and thus a fair model, is to minimize discrimination by the algorithm. Important to note, however, Feuerriegel et al. [19] states that fair AI algorithms can have different objectives. Firstly, the goal of an algorithm can be to assess its fairness by means of measuring the prediction performance. This can, for example, be done by looking at the type-I and type-II errors across subgroups. Secondly, the goal of a fair algorithm can be to produce fair forecasts. For this, approaches such as pre-processing data, adjusting underlying classifiers and post-processing data are used [20]. Lastly, a possible goal can be to model fair decision making tasks, which typically need customised systems that precisely model feedback loops.

2.3 Fairness definitions and measures

The translation from fairness in society to fairness in models is not as easy as it seems. Fairness aims to minimize the discrimination in a model, whether it is positive or negative, but how fairness is defined and measured is different in many of the literature. Verma and Rubin [10] reviewed multiple machine learning articles and papers and consider twenty fairness definitions related to predictive machine learning models. According to Naraynan, the number of definitions is "non-exhaustive," making it difficult to choose a single term and thus approach [21].

For example, scholars state the following definition: balance for a class requires that the average score assigned to individuals of group A who belong to that class, should be the same as the average score assigned to individuals of group B who belong to that class [22]. In other words, the assignment of scores should not be systematically more inaccurate for instances of the same class in one group than the other. Moreover, Kasy and Abebe [23] discuss fairness as the notion of 'merit', denoting that groups of equal 'merit' should be treated in the same manner.

Huang et al. [14] have an extensive definitions list of different types of fairness which is cited often, and theirs will be used as base definitions in this research. Verma and Rubin list the mathematical functions corresponding to these fairness definitions [10]. The comparison of the true positive, true negative, false positive and false negative rates, stated as TPR, TNR, FPR and FNR respectively, is made between the protected group versus the non-protected group. For consistent purposes, the protected group will be stated as female versus the non-protected group; male. Moreover, the positive rate means the rate of being selected by the algorithm.

Huang et al. discuss three types of fairness:

- **Unawareness**: this is a type of fairness where the model does not have any of the sensitive variables as part of the input. Therefore they play no role on the decision process. However, this raises the question of how the model's fairness can be measured if the sensitive variables are removed [24]. Additionally, when deleting these sensitive variables, the algorithm can still track statistically significant patterns form the data that is left over. It may discover characteristics that are associated to specific category data (proxies) and employ these proxies when training the decision model [25]. For example, when an algorithm finds correlation between gender and another non-protected variable. This can indirectly correlate to the target value.

- **Individual fairness**: this type of fairness has as goal that 'similar people are treated similarly' [26]. Meaning that any pair of individuals should be treated equally.
- **Group fairness**: this type of fairness is split up into six equations to measure fairness, where the statistics should be similar across groups.
    - *Statistical parity* - Statistical parity is the notion that equal fractions of each group should be treated as belonging to the positive class [27] [28] [29]. This can be used such that algorithms penalize statistical parity in their modelling [27] [29].
      The respective equation for this definition is:
      $$PositiveRate(Male) = PositiveRate(Female) \tag{1}$$
    - *Equal opportunity* - Equal opportunity means that any individual who is qualified should have the similar opportunities for the same outcome [30].
      The respective equation for this definition is:
      $$TPR(Male) = TPR(Female) \tag{2}$$
    - *Equalized odds* - Similar to equal opportunity - This fairness definition requires there to be equal acceptance rate (the number of individuals who pass should be the same as the number of individuals who do not pass in both groups) [30].
      The respective equations for this definition is equation (2) and:
      $$FPR(Male) = FPR(Female) \tag{3}$$
    - *False positive rate balance* - This fairness definition focuses on equal false positive rates of both protected and unprotected group.
      The respective equation for this definition is equation (3).
    - *False negative rate balance* - This fairness definition focuses on equal false negative rates of both protected and unprotected group.
      The respective equation for this definition is:
      $$FNR(Male) = FNR(Female) \tag{4}$$
    - *Equal treatment* - This definition focuses on the error rate of the model. The error rate, so the false positives and false negatives, for both protected and unprotected groups should be equal.
      The respective equations for this definition is equation (3) and (4).
    - *Overall accuracy equality* - This fairness definitions focuses on equal prediction accuracy of both protected and unprotected groups.
      The respective equation for this definition is:
      $$Accuracy(Male) = Accuracy(Female) \tag{5}$$
- **Counterfactual fairness**: this is a type of fairness where the decision or prediction regarding an individual should not change if the environment or value of the sensitive variable changes.

2.4 Bias in AI

As mentioned before, bias refers to a system or model that discriminates a person or a group of people based on a feature [18]. Bias can occur in data and in an AI model [31]. There are many causes of bias in dataset, i.e. historical biases, imbalance, etc. Algorithms are in turn biased as they are trained on such datasets. To improve fairness of predictive models, the literature addresses detecting bias, identifying bias sources, and mitigating algorithmic bias [32]. Unfairness can be caused by, for example, sample bias, historical biases, or ingenuous training of the model with extreme class-imbalance [33] [34]. Fairness is measured after a prediction model is made, whereas bias can already be measured in the dataset. As Saleiro et al. measure bias and fairness is both at the end of the process [35]. The absence of bias in a dataset or in a model will result in a fair model.

2.5 Techniques for reducing unfairness and bias

Techniques for reducing unfairness and bias are typically categorized into three techniques, namely pre-processing, during processing, and post-processing [36].

- Pre-processing. Here the training data gets adapted before using it as input for predictive modelling. Variables are adapted such that discriminating relationships are altered within the data. These variables could be sensitive variables. This method includes unawareness and counterfactual fairness approached as for both attributes should be deleted or changed beforehand.
- In-processing. The machine learning algorithms are altered in such a way that it takes fairness into account [37] [38]. This could include, for example, penalization of incorrect classifications or adding constraints to a machine learning model.
- Post-processing. The outputs of prediction models are used in different ways to increase fairness. An example is to flip some decisions of a prediction model to increase equalized odds and equalized opportunity [30]. Another option would be to change thresholds for different sensitive groups upon decision making [39] [40].

3 Methodology

Following that, the methods of this study will be described. Because the goal of this research is to assist real-world scenarios and issues related to fairness, the model will be developed in a practical manner.

3.1 Research Strategy

Three research strategies that aim to answer the research sub questions, and therefor the main research question, are specified. The following approach is conducted: qualitative descriptive research, implementation on the model, and a field survey.

Qualitative descriptive research. To answer the first sub research question, a qualitative descriptive research is conducted. A thorough literature review of articles regarding fairness definitions and its appliance to fairness in AI is conducted. The definitions should be reviewed with respect to the chosen recruitment dataset to have only meaningful, relevant fairness definitions left. Furthermore, the metrics of fairness and bias in a recruitment dataset are investigated as well as the techniques for reducing unfairness and bias relating to the reviewed fairness definitions.

Data analysis and processing in Python. I will make a recruitment prediction model which should have a good accuracy level. Then fairness values corresponding to the different fairness definitions are calculated. Based on these outcomes the prediction model will be enhanced according to the reducing of unfairness and bias technique(s). Sub research question four is answered within this part.

Field survey. To evaluate whether the fairness outcomes of the reduced unfairness and bias recruitment prediction model is agreed by the fairness perceptions of society, we will conduct a field survey. As different societal viewpoints want to be examined, the survey is proposed to different groups, namely potential applicants, recruiters and AI experts. A recruiting process has impact on potential applicants. It is hosted by recruiters. And whenever an AI prediction model is build for the recruitment process, AI experts are involved as well. Here the survey participants are asked to rank the application of fairness definitions within a recruitment model. This analysis will identify weak fairness definitions with respect to the recruitment use-case and will give guidance to future development of fair algorithms.

3.2 Dataset

The dataset is selected upon a few criteria, namely:

- A predictable variable related to recruitment where people are listed with their regarding target variable; recruited or not recruited,
- Enough entries for prediction, and
- A minimum of one sensitive variable, where gender variable is preferred.

Based on these criteria datasets were found via web search. This led to the conclusion that there only a handful of datasets satisfying the first criteria and only three which satisfy all criteria points. These are found on Kaggle and Github.
The first relevant dataset from Rohan Lal Kshetry and gathered from Kaggle [41] has 614 entries. Further details of the data can be found in Appendix A.
The second relevant dataset is from the People Analytics Project, created by Sambit Das, and gathered from Github [42]. The dataset has 280 entries. This

dataset has less entries than Kshetry's dataset. Further details of the data can be found in Appendix B.

The third relevant dataset is created by C.U. Khamaleswar and is gathered from Kaggle [43]. The dataset has 9528 entries. However, the correlation between the attributes and the target variable was too low for a well-predicting model. Further details of the data can be found in Appendix C.

Kshetry's (2021) [41] dataset on Indian CV hiring was most suitable as it has a reasonable size, includes the sensitive variable gender and the variables are correlated to the target variables, which is needed for predictive modelling. The dataset contains 10 attributes as indicated in the figure below.

Attribute	Description
ID	Unique identifier for each candidate
Gender	Sex of the candidate
Python experience	Boolean whether the candidate has previous Python experience
Years of experience	Amount of years the candidate has worked before
Education	Boolean whether the candidate is graduated or not
Internship	Boolean whether the candidate has followed an internship
Score	Numerical value
Salary	Amount of money the candidate is earning currently
Offer history	Boolean whether candidate has got offered a job before
Location	Residence of the candidate (Rural, suburban, urban)
Recruitment status	Boolean whether candidate is recruited or not

Table 1: Recruitment dataset characteristics

3.3 Performance measure

The machine learning model is assessed based on the accuracy, precision and recall of the model. This are a widely used measures for evaluation of a model's performance.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

Furthermore, F1 score is the harmonic mean of precision and recall. It can be useful for evaluation measures when a dataset is imbalanced [44]. However, the recruitment dataset used is balanced enough such that accuracy will fulfill for means of evaluation of the model's performance.

3.4 Fairness value

Section 2.3 states the different fairness definitions and their respective equations to measure these definitions. These are calculated by means of the variables within the confusion matrix. These are split up into:

- True positive rate. Where true positive means the candidates that are predicted to be recruited and according to the data are also recruited.
- True negative rate. Where true negative means the candidates that are predicted to not be recruited and according to the data are also not recruited.
- False positive rate. Where false positive means the candidates that are predicted to be recruited and according to the data are, however, not recruited.
- False negative rate. Where false negative means the candidates that are predicted to not be recruited and according to the data are, however, recruited.

Whenever these values are calculated. A ratio of the measures for both protected and unprotected group can be calculated. This will be stated as the fairness value of that respective definition. The fairness values of each fairness definition are listed below.
Statistical parity:
For clarity, Positive Rate is abbreviated to PR.

$$fairness\_value = \begin{cases} \frac{PR(Male)}{PR(Female)} & \text{if } PR(Male) < PR(Female) \\ \frac{PR(Female)}{PR(Male)} & \text{if } PR(Female) < PR(Male), \\ 1, & \text{if } PR(Male) = PR(Female) \end{cases} \tag{9}$$

Equal opportunity:

$$fairness\_value = \begin{cases} \frac{TPR(Male)}{TPR(Female)} & \text{if } TPR(Male) < TPR(Female) \\ \frac{TPR(Female)}{TPR(Male)} & \text{if } TPR(Female) < TPR(Male), \\ 1, & \text{if } TPR(Male) = TPR(Female) \end{cases} \tag{10}$$

Equalized odds:

$$TPR\_value = \begin{cases} \frac{TPR(Male)}{TPR(Female)} & \text{if } TPR(Male) < TPR(Female) \\ \frac{TPR(Female)}{TPR(Male)} & \text{if } TPR(Female) < TPR(Male), \\ 1, & \text{if } TPR(Male) = TPR(Female) \end{cases} \tag{11}$$

$$FPR\_value = \begin{cases} \frac{FPR(Male)}{FPR(Female)} & \text{if } FPR(Male) < FPR(Female) \\ \frac{FPR(Female)}{FPR(Male)} & \text{if } FPR(Female) < FPR(Male), \\ 1, & \text{if } FPR(Male) = FPR(Female) \end{cases} \tag{12}$$

$$fairness\_value = \frac{TPR\_value + FPR\_value}{2} \tag{13}$$

False positive rate balance:

$$fairness\_value = \begin{cases} \frac{FPR(Male)}{FPR(Female)} & \text{if } FPR(Male) < FPR(Female) \\ \frac{FPR(Female)}{FPR(Male)} & \text{if } FPR(Female) < FPR(Male), \\ 1, & \text{if } FPR(Male) = FPR(Female) \end{cases} \tag{14}$$

False negative rate balance:

$$fairness\_value = \begin{cases} \frac{FNR(Male)}{FNR(Female)} & \text{if } FNR(Male) < FNR(Female) \\ \frac{FNR(Female)}{FNR(Male)} & \text{if } FNR(Female) < FNR(Male), \\ 1, & \text{if } FNR(Male) = FNR(Female) \end{cases} \tag{15}$$

Equal treatment:

$$FPR\_value = \begin{cases} \frac{FPR(Male)}{FPR(Female)} & \text{if } FPR(Male) < FPR(Female) \\ \frac{FPR(Female)}{FPR(Male)} & \text{if } FPR(Female) < FPR(Male), \\ 1, & \text{if } FPR(Male) = FPR(Female) \end{cases} \tag{16}$$

$$FNR\_value = \begin{cases} \frac{FNR(Male)}{FNR(Female)} & \text{if } FNR(Male) < FNR(Female) \\ \frac{FNR(Female)}{FNR(Male)} & \text{if } FNR(Female) < FNR(Male), \\ 1, & \text{if } FNR(Male) = FNR(Female) \end{cases} \tag{17}$$

$$fairness\_value = \frac{FPR\_value + FNR\_value}{2} \tag{18}$$

Overall accuracy equality:

$$fairness\_value = \begin{cases} \frac{Accuracy(Male)}{Accuracy(Female)} & \text{if } Accuracy(Male) < Accuracy(Female) \\ \frac{Accuracy(Female)}{Accuracy(Male)} & \text{if } Accuracy(Female) < Accuracy(Male), \\ 1, & \text{if } Accuracy(Male) = Accuracy(Female) \end{cases} \tag{19}$$

For each of these definitions, the fairness value will lead to a value between 0 and 1. Where 0 means that the model, given the set parameters, is completely unfair according to the regarding fairness definition, and vice versa for 1; meaning a fair model.

4 Analysis and results

This section describes the data that is used for the predictive modeling use-case. Moreover, it states the models that are trained, and the results given the field survey.

4.1 Data pre-processing and exploration

For data pre-processing and exploration, Python 3.8 was used in this project. It offers a wide range of data analytics tools with pre-coded libraries. A first step examining Kshetry's [41] dataset was to look for missing values. Only the Score, Location and Recruitment status attributes did not miss any values. The variable with the most missing entries was the Internship variable with 32 missing entries. The missing values were chosen to be replaced with the most occurring option for categorical variables and the mean for numerical variables.
For clarity, offer history does not mean that a person is offered this job because there are some rows where the person is recruited but does not have an offer history.// The next step was to make binary variables out of each categorical variable. Meaning that, for example, the Location attribute is split into Rural, Urban and Semiurban having 0 or 1 as entry values.
Furthermore, each numerical variable is normalized within a 0 to 1 range. This makes it easier to read and used in the same way across the dataset.
The resulting dataset is described in the table below.

	Male	Female	Rural	Urban	Semiurban	Graduate	Not_Graduate	Internship
count	613	613	613	613	613	613	613	613
mean	0.817	0.182	0.292	0.327	0.380	0.781	0.218	0.133
std	0.386	0.386	0.455	0.469	0.485	0.413	0.413	0.340
min	0	0	0	0	0	0	0	0
25%	1	0	0	0	0	1	0	0
50%	1	0	0	0	0	1	0	0
75%	1	0	1	1	1	1	0	0
max	1	1	1	1	1	1	1	1

	Python_exp	Exp_years	Score	Salary	Offer_hist	Recruitment_status
count	613	613	613	613	613	613
mean	0.345	0.769	0.257	0.065	0.854	0.687
std	0.476	0.334	0.076	0.121	0.352	0.464
min	0	0	0	0	0	0
25%	0	0	0.033	0.131	1	0
50%	0	0	0.045	0.173	1	1
75%	1	0.333	0.069	0.225	1	1
max	1	1	1	1	1	1

Table 2: Feature characteristics of recruitment dataset

The tables above describe the preprocessed dataset. The imbalance in males and females is remarkable. The split of recruiting females versus males is also imbalanced as can be seen in the figure below. The dataset is imbalanced towards having more men then women, see the bottom row of figure 1, and having more recruited people than non-recruited people, see the last column of figure 1. Most of the data points are recruited men.

Recruitment status is the target value of this data set. Each attribute is has a correlation to this target attribute. This correlation can be viewed on the bottom line of figure 7. Here, offer history has the highest correlation to whether a person is recruited or not. Living locations such as semiurban and rural also have a one of the highest correlations.

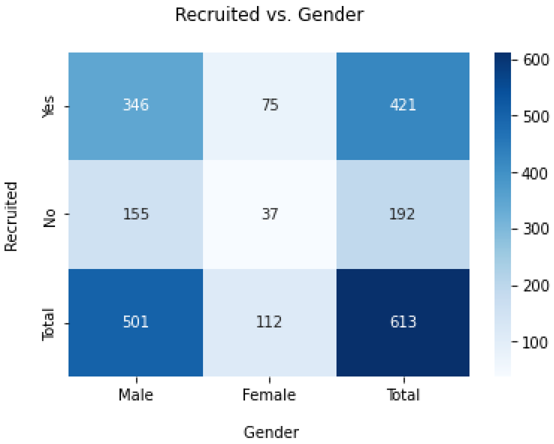


Fig. 1: Gender vs. Recruited in dataset.

4.2 Model training

Before a model can be trained and tested appropriately, the dataset split into a train and test set by means of cross validation. Python's scikit-learn built-in train-test split function is used to split the dataset in an 80-20 ratio while taking a random sample. A random-state of 42 is included in the function for replicating purposes.

As a next step, this research considers multiple models for recruitment prediction. For prediction models to be used, the accuracy should be as high as possible. As the dataset has more recruited persons than non-recruited persons, a baseline that can be used is a model where everyone is recruited. This model will be called the 'Dum' model as it does not reason behind it's decision of recruitment. Furthermore, Multilayer Perceptrons (MLP), XGBoost, Random Forest and Logistic Regression are used for predictive modelling. Each model will generate a probability of recruiting a specific person within the train and test dataset.

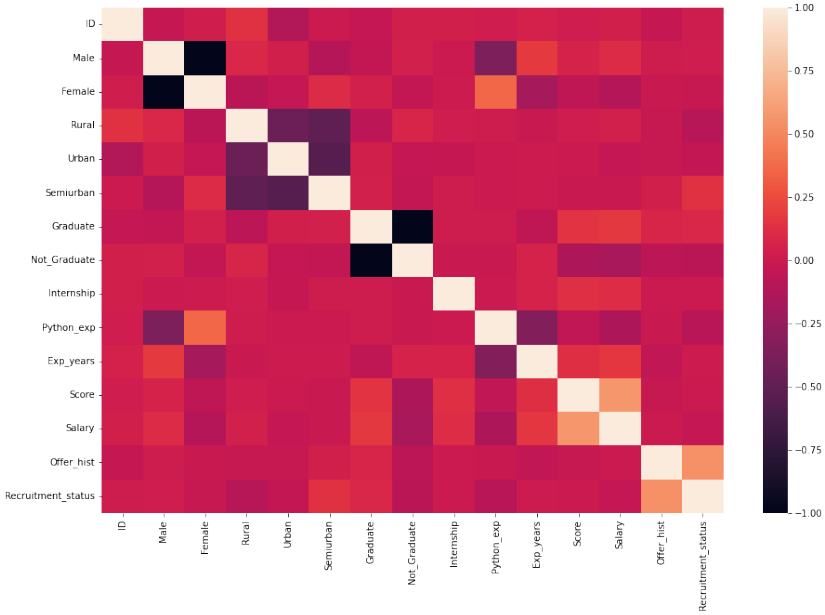


Fig. 2: Correlation matrix.

From these results, the accuracies of each of the models can be calculated and are described below.

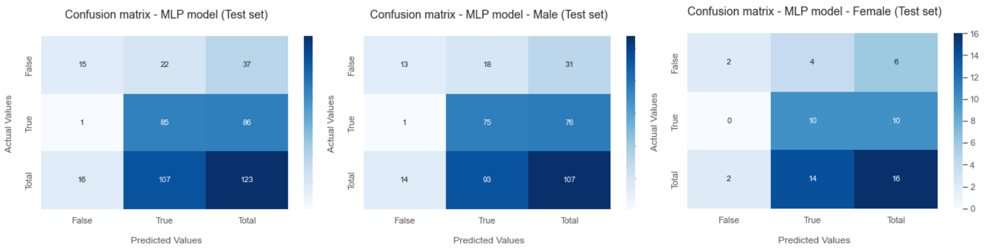
Having the Dum model figure as a baseline model and aiming to have the highest accuracy, the MLP will feature as the standard model for the succeeding part of this project.

Dataset	Dum	MLP	XGBoost	Random Forest	Logistic Regression
Train	0.684	0.808	0.494	0.727	0.799
Test	0.699	0.813	0.494	0.775	0.801

Table 3: Accuracies of prediction models

MLP equal thresholds. As the MLP model has the highest accuracy of the different prediction models, it will be used as a base model for the rest of the project. This MLP model wil use equal threshold for male and female, therefore it is referred to as MLP_ET. Using this model for finding the persons to recruit, the fairness values can be measured for the predictions. The predictions of the MLP model can be viewed in figure 3 which visualises the prediction on the test

set and the predictions split up in the male and female candidates. For these predictions the thresholds for both males and females are set on 0.5. Meaning that a probability, generated by the MLP model, higher than 0.5 would recruit the candidate, see appendix G MLP_ET.
As discussed in the methodology section, changing thresholds of variables will be used for enhancing the model with respect to the fairness definitions. Whenever the thresholds for both male and female candidates are remained equal but changed in parallel to increase the fairness values, this will give the following results, see table 4.



(a) Confusion matrix - MLP model.

(b) Confusion matrix - MLP model - Male.

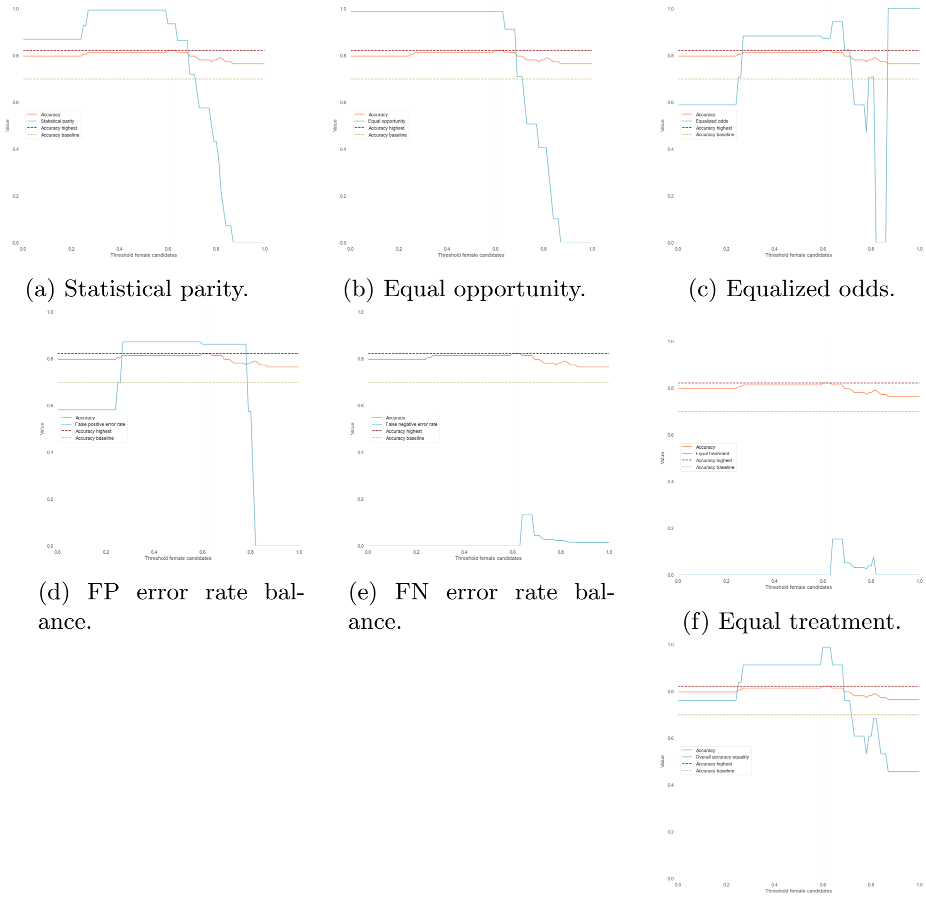
(c) Confusion matrix - MLP model - Female.

Fig. 3: Confusion matrices of the MLP model

Threshold range/	Highest accuracy		Accuracy above baseline		Total	
Fairness measure	Value	Threshold	Value	Threshold	Value	Threshold
Statistical parity	0.935	0.60	0.993	0.27	0.993	0.27
Equal opportunity	0.987	0.60	0.987	0	0.987	0
Equalized odds	0.873	0.60	1	0.87	1	0.87
FP error rate balance	0.861	0.60	0.871	0.27	0.871	0.27
FN error rate balance	0	0.60	0.132	0.64	0.132	0.64
Equal treatment	0	0.60	0.153	0.64	0.153	0.64
Equal accuracy	0.988	0.60	0.988	0.60	0.988	0.60

Table 4: Fairness measures split up into multiple accuracy scenarios when keeping thresholds for male and female candidates equal

MLP different thresholds. To see whether the MLP_ET can enhanced the threshold of the protected group, females, are changed whenever the threshold of males are kept at 0.5. This model is referred to as MLP_DT. This is done because whenever having for example a higher chance for females of being recruited, meaning the threshold is lower, this will maybe result in a more fair model.



(a) Statistical parity.

(b) Equal opportunity.

(c) Equalized odds.

(d) FP error rate balance.

(e) FN error rate balance.

(f) Equal treatment.

(g) Equal accuracy.

Fig. 4: Fairness values for each female candidate threshold

Fairness measure	Fairness value	Female threshold	Accuracy model
Statistical parity	0.955	0.5	0.806
Equal opportunity	0.987	0.6	0.813
Equalized odds	1	0.9	0.733
FP error rate balance	0.842	0.5	0.806
FN error rate balance	0.164	0.63	0.806
Equal treatment	0.183	0.63	0.806
Equal accuracy	0.987	0.6	0.813

Table 5: For each fairness measure, the female threshold value is shown that optimises fairness. Next to it the resulting accuracy is shown.

As can be seen the fairness values improved a bit when changing the threshold of the female candidates when holding the threshold of the male candidates equal.

Ranking of fairness measures The findings mentioned above will lead to a ranking of fairness within the enhanced MLP_DT model. Based on giving highest meaning to having a model that performs more accurately than the benchmark, but having a goal to have the highest fairness value will give the following ranking.

Rank	Fairness measure
1	Equalized odds
2	Equal accuracy
3	Equal opportunity
4	Statistical parity
5	FP error rate balance
6	Equal treatment
7	FN error rate balance

Table 6: Ranking of fairness measures regarding the results of the MLP_DT model

4.3 Survey

Population The survey is presented to three different societal groups that are involved within recruitment process collaborative with AI model, namely recruiters, potential applicants and AI experts. The distribution of the participants is shown below.

Ranking of fairness definition Each survey participant, from either the recruiters, potential applicants or AI experts group, are asked to give their opinion on the level of fairness within the sketched scenario. A scale-based answer method is used, where giving a “1” means the scenario is completely unfair, whereas “6” means a completely fair scenario.

Legitimate factors The survey participants were also asked to consider which factors are important for recruitment. Here they could choose one or multiple factors. A rank is made which factors are most important to least important, see the table below. These factors could be used for further development of models

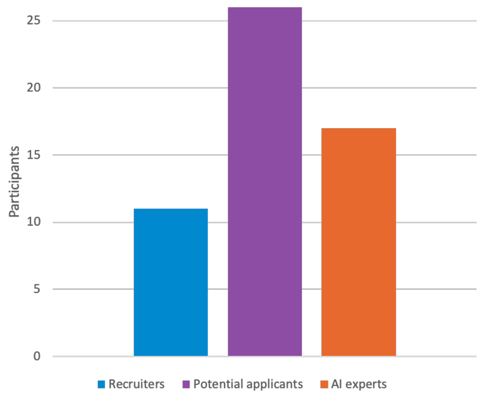


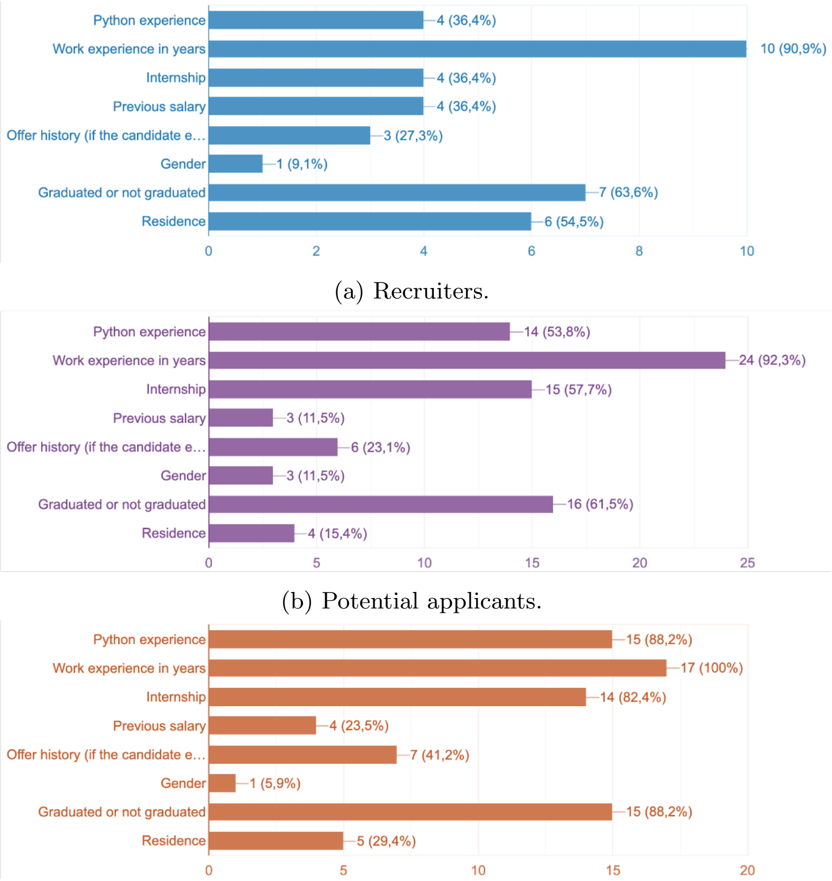
Fig. 5: Distribution of the survey participants.

Rank	Fairness measure	Value (0-5)
1	Individual fairness	4.13
2	False positive rate balance	3.93
3	False negative rate balance	3.93
4	Equal treatment	3.93
5	Predictive parity	3.13
6	Unawareness	3.12
7	Counterfactual fairness	2.89
8	Overall accuracy equality	2.69
9	Statistical parity	2.54
10	Equal opportunity	2.35
11	Equalized odds	2.35

Table 7: Ranking of fairness measures regarding survey results

Rank	Legitimate factor (all)	Legitimate factors (recruiters)
1	Years of experience	Years of experience
2	Education	Education
3	Python experience	Location
4	Internship	Python experience
5	Location	Internship
6	Offer history	Salary
7	Salary	Offer history
8	Gender	Gender

Table 8: Ranking of legitimate factors



(a) Recruiters.

(b) Potential applicants.

(c) AI experts.

Fig. 6: Answers regarding legitimate factors question

including the most important factors.
Interesting here is that “gender” is classified as a non-important attribute for candidate selection. The rank-list is more or less the same for the total ranking as for the recruiters in particular, except for location.

Further findings An interesting finding regards the preference of whether the recruitment process should only involves humans or AI collaboration. The distribution is shown below. As can be seen in the figure above, potential applicants prefer the involvement of only humans in the recruitment process. Whereas, the majority of recruiters would like a collaboration with AI. When looking into the participant data, it can be seen that the majority of recruiting wanting AI collaboration is in the top percentile of recruiting experience.

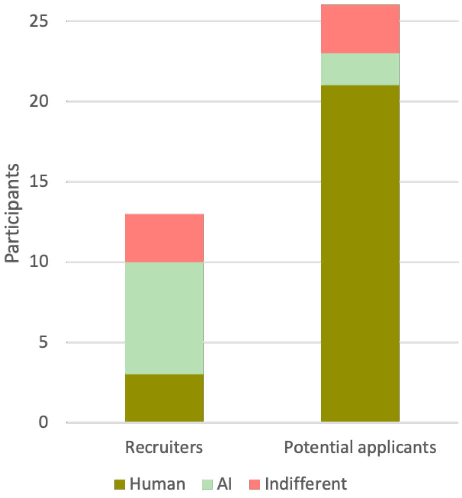


Fig. 7: Distribution of preferences.

From the survey answers, there were two questions where the answers sets were very divided among the the three participant groups. The opinion is measured whether applying the fairness definitions will mean that the model will be fair. The first one is regarding unawareness and the second is regarding overall accuracy equality. The distribution of answers regarding the unawareness question is shown in figure 8. The majority of recruiters find that an unaware model will be fair. Whereas potential applicants lean more towards unfair, and AI experts also lean a bit towards unfair. The second regards the overall accuracy equality questions which can be found in figure 9. Recruiters and potential applicants find the appliance of the overall accuracy equality not meaning that the model will not be discriminatory. Whereas AI experts are very divided whether this definition will result in a fair model.

Other interesting findings regard the open survey questions regarding the testing of a recruitment AI model. These questions were asked to recruiters and AI experts, to see the different viewpoints of these groups. Most recruiters would test recruitment software before implementing it into their company even though an external organisation tested it to be 99% accurate. The recruiter participants proposed an approach of how they would test the model.

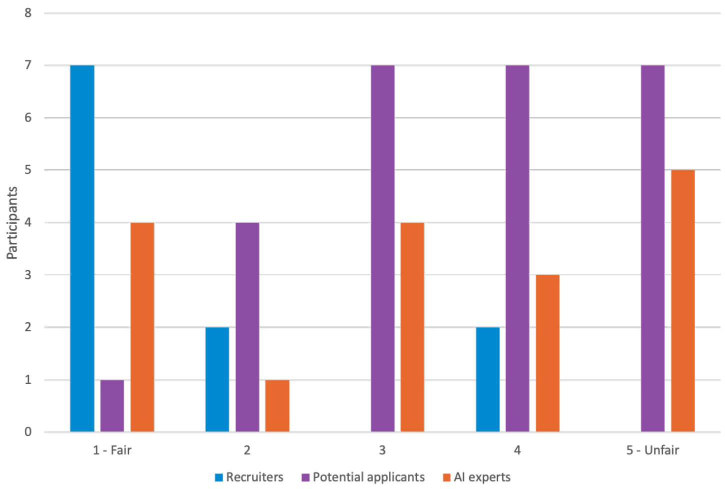


Fig. 8: Distribution of survey answers regarding the question about unawareness.

1. Evaluation. Almost all recruiters would use evaluation as a method of testing a recruitment model. They would compare the results of the model with their own findings on who to recruit.
2. Reasoning. A few recruiters would find it important to know the choices the algorithm made and why these choices are made. The reasoning behind the model can be used as a way to see whether the model is acceptable.
3. Dummy test. One recruiter would also use the evaluation method, but with dummy persons instead of real candidates.
4. Hand over. One recruiter would hand over the testing to an internal developer and would not interfere with the process anymore.

Subsequently, most recruiters would not test the AI model for themselves if it is already tested by their own company and tested 99% accurate.

As for AI experts, they agree in almost all cases upon the test approach. The approaches can be split up into the following:

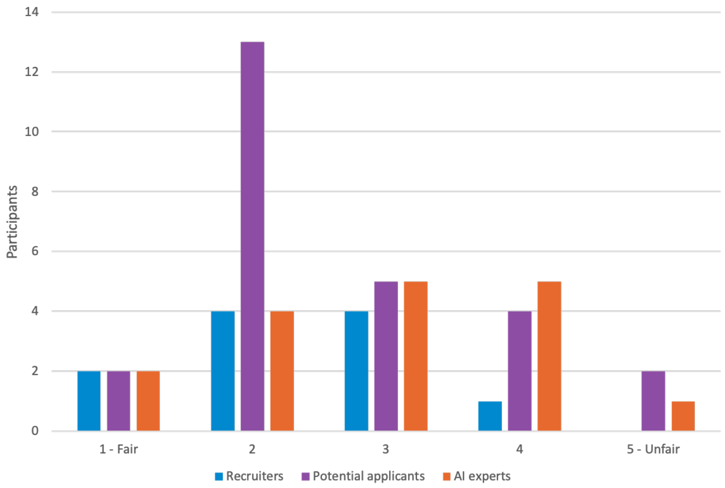


Fig. 9: Distribution of answers regarding the question about overall accuracy equality.

1. Gather CV data, label the data by professionals/recruiters and evaluate the algorithm.
2. Same as (1), but removing sensitive variables.
3. Same as (1), but focusing on fairness measures when evaluating.
4. Same as (1), but also evaluate after one year.
5. Same as (1), but also test on reasonability of decisions and robustness.

Furthermore, AI experts would test the algorithm differently for each company. Where they also add to take into account the company's sensitive variables, standards for a good CV and standards for a position or role.

When comparing recruiters and AI experts on their testing viewpoints, recruiters focus more on the evaluation of their own outcomes, where AI recruiters look into it a few steps deeper. Meaning they also consider sensitive variables fairness measures, and robustness.

5 Discussion and conclusion

This research has shown that there are many different definitions for fairness. A selection of a few definitions is made to be used as evaluation of a recruitment prediction model. Fairness types are split into unawareness, individual fairness and group fairness. Statistical parity, equal opportunity, equalized odds, false positive rate balance, false negative rate balance, equal treatment and overall accuracy are considered for evaluation purposes. Each fairness definition has a correlating mathematical expression. This is combined into a fairness value,

upon which the recruitment model is evaluated.
Techniques for reducing unfairness and bias include pre-processing and post-processing options. Pre-processing is linked to unawareness and counterfactual fairness definitions. Whereas, post-processing can be used for each fairness definitions. Here the threshold for recruiting male and female candidates is altered. When setting the male threshold to a standard of recruiting candidates whenever the MLP probability is higher than 0.5, and changing the female threshold between 0 and 1, this will state that equalized odds can be the most fair and false negative rate balance the least for this recruitment case. However, two of the fairness values remain very small, meaning that the model cannot be indicated as fair regarding these definitions. The main plausible reason for this is that the dataset is very imbalanced. Most of the candidates are recruited and men. This means that the MLP model cannot properly train.
Besides predictive modelling, society was asked upon their opinion about using collaborative AI models for the help of recruitment processes. Individual fairness is by far regarded as the most fair to base a fairness definition on. Whereas, equalized odds and equal opportunity are seen as the least fair approaches. The opinions on unawareness and overall accuracy equality definitions were very divided. Where recruiters and and AI experts were most divided in their opinions on fairness. Moreover, the field survey found that years of work experience, previous education and previous python experience are seen by these participants as important factors to base a recruitment decision on. Furthermore, the approaches of testing a prediction model before implementing it into society is viewed differently. Recruiters focus more on their own comparison of results, whereas AI experts find it more interesting to look at the approach of comparison; where to look at, what to include and remaining the same within a year.
All in all, fairness within a recruitment prediction model can be measured and improved in various ways. This will not always imply that each fairness value can be optimized to a fair model due to poor datasets. Moreover, society is still indecisive about the various fairness definitions and how an AI model should behave and should be used.

Theoretical contribution

This research gives a overview of the various fairness definitions, their mathematical calculations, and how to calculate fairness values accordingly. An interpretation of these fairness values are given. Moreover, this is all applied to a real-world use-case about prediction of recruitment. The results give an indication of which fairness definitions work best for this recruitment use-case. Furthermore, the field survey gives insights on the interpretation of fairness in society groups involved in recruitment processes; recruiters, potential applicants and AI experts.

Practical contribution

This project gives a practical contribution regarding the code that is created to built a prediction model for the recruitment use-case. This code can be used as a guideline on how to pre-process recruitment data, built a prediction model, calculate fairness values, compare fairness values with accuracy, give insightful figures on how the model is fair, and how to post-process the model to enhance it according to fairness. Furthermore, a structured overview of different questions for survey purposes regarding fairness definitions is given. These can be used for further research.

Limitations and future work

The project was based on the Indian recruitment dataset of Kshetry [41]. For Dutch purposes, it would be more useful when the use-case is based on a dataset with Dutch persons in it. Additionally, the dataset is imbalanced. A balanced dataset would perhaps improve the fairness of a model. Whenever a new dataset is created, include as much sensitive variables and entries as possible. The more sensitive variables you have the more the model can be checked whether it is not discriminating. More entries will lead to more involved different cases for better results.

The dataset appeared to be more imbalanced as foreseen before. For future work with this recruitment dataset, it is encouraged to also look at the F1-score to evaluate the performance of the model.

For more completeness, it would be perhaps interesting to also look at other fairness cases. These could be fairness definitions regarding fairness in relational domains, fair inference, well calibration, etc. [10]

Moreover, Google's What-If Tool [45] and IBM's AI Fairness 360 [46] can also be used to evaluate fairness of models. In essence, these libraries help with testing the performance in fictitious scenarios, evaluate the significance of various data attributes and show model behavior across many models and subsets of input data.

According to the results of the survey, legitimate factors that should be considered are years of work experience, previous education and previous python experience. A new model can be trained based on these legitimate factors. These attributes can be included for the training of future models as they appear to be important to the three different society groups.

References

1. J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women." 2018.
2. J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency.* PMLR, 2018, pp. 77–91.

3. A. International, “Toeslagenschandaal is mensenrechtenschending, zegt amnesty international,” 2021.
4. S. Wachter, “Current discrimination laws failing to protect people from ai-generated unfair outcomes,” *Oxford Internet Institute*. [Online]. Available: https://www.oii.ox.ac.uk/news-events/news/ai-creates-unintuitive-and-unconventional-groups-to-make-life-changing-decisions-yet-current-laws-do-not-protect-group-members-from-ai-generated-unfair-outcomes-says-new-paper/
5. P. Crosman, “Unfair lending with ai? don't point just at us, fintechs and online lenders say,” *American Banker*. [Online]. Available: https://www.americanbanker.com/news/unfair-lending-with-ai-dont-point-just-at-us-fintechs-and-online-lenders-say
6. “Ai at work: Staff 'hired and fired by algorithm',” *BBC News*. [Online]. Available: https://www.bbc.com/news/technology-56515827
7. L. Sweeney, “Discrimination in online ad delivery,” *Communications of the ACM*, vol. 56, no. 5, pp. 44–54, 2013.
8. S. V. Otterloo, “Algorithmic bias and how to avoid it in software projects,” 2018. [Online]. Available: https://ictinstitute.nl/algorithmic-bias-project- management/
9. D. Alvarez Melis and T. Jaakkola, “Towards robust interpretability with self-explaining neural networks,” *Advances in neural information processing systems*, vol. 31, 2018.
10. S. Verma and J. Rubin, “Fairness definitions explained,” in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.
11. B. Hutchinson and M. Mitchell, “50 years of test (un) fairness: Lessons for machine learning,” in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 49–58.
12. P. Snel and S. van Otterloo, “Practical bias correction in neural networks: a credit default prediction case study,” 2022.
13. “The oecd artificial intelligence (ai) principles - oecd.ai,” https://oecd.ai/en/ai-principles, accessed: 2022-07-27.
14. R. Fu, Y. Huang, and P. V. Singh, “Artificial intelligence and algorithmic bias: Source, detection, mitigation, and implications,” in *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*. INFORMS, 2020, pp. 39–63.
15. M. F. Özbilgin and C. Erbil, “Social movements and wellbeing in organizations from multilevel and intersectional perspectives: The case of the# blacklivesmatter movement,” *The SAGE Handbook of Organizational Wellbeing*, pp. 119–138, 2021.
16. “Artikel 1 burgerlijk wetboek,” in *De Grondwet*. Nederlandse Overheid. [Online]. Available: https://wetten.overheid.nl/
17. O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, “State-of-the-art in artificial neural network applications: A survey,” *Heliyon*, vol. 4, no. 11, p. e00938, 2018.
18. B. Friedman and H. Nissenbaum, “Bias in computer systems,” *ACM Transactions on Information Systems (TOIS)*, vol. 14, no. 3, pp. 330–347, 1996.
19. S. Feuerriegel, M. Dolata, and G. Schwabe, “Fair ai,” *Business & information systems engineering*, vol. 62, no. 4, pp. 379–384, 2020.
20. S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 329–338.
21. A. Narayanan, “Translation tutorial: 21 fairness definitions and their politics,” in *Proc. Conf. Fairness Accountability Transp., New York, USA*, vol. 1170, 2018, p. 3.

22. J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.
23. M. Kasy and R. Abebe, "Fairness, equality, and power in algorithmic decision-making," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 576–586.
24. T. Slaff, R. Gieske, Sharon, H. Casper, Van Havendonk, E. Webbe, R. Van den Akker, G. Sileno, and C. Bruno, "Ethical ai: how to lead?" *WHITEPAPER AMSTERDAM DATA COLLECTIVE*, 2021.
25. S. Ruggieri, D. Pedreschi, and F. Turini, "Data mining for discrimination discovery," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 4, no. 2, pp. 1–40, 2010.
26. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.
27. T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data mining and knowledge discovery*, vol. 21, no. 2, pp. 277–292, 2010.
28. F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd international conference on computer, control and communication.* IEEE, 2009, pp. 1–6.
29. T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *2011 IEEE 11th International Conference on Data Mining Workshops.* IEEE, 2011, pp. 643–650.
30. M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.
31. J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein, "Discrimination in the age of algorithms," *Journal of Legal Analysis*, vol. 10, pp. 113–174, 2018.
32. A. M. F. da Cruz, "Fairness-aware hyperparameter optimization," 2020.
33. H. He and Y. Ma, "Imbalanced learning: foundations, algorithms, and applications," 2013.
34. H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," *arXiv preprint arXiv:1901.10002*, vol. 2, 2019.
35. P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.
36. J. Silberg and J. Manyika, "Notes from the ai frontier: Tackling bias in ai (and in humans)," *McKinsey Global Institute*, pp. 1–6, 2019.
37. A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning.* PMLR, 2018, pp. 60–69.
38. Y. Bechavod and K. Ligett, "Penalizing unfairness in binary classification," *arXiv preprint arXiv:1707.00044*, 2017.
39. S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.
40. A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Conference on Fairness, Accountability and Transparency.* PMLR, 2018, pp. 107–118.
41. R. L. Kshetry, "Should recruit or not," 2021. [Online]. Available: https://www.kaggle.com/code/rafunlearnhub/should-recruit-or-not/data

42. S. Das, "People analytics project," 2019. [Online]. Available: https://github.com/Sambit78/People-Analytics-Project
43. C. Khamaleswar, "Av hiring," 2020. [Online]. Available: https://www.kaggle.com/datasets/khamalking/avhiring?select=train$_K NurWLh.csv$
44. Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize f1 score," *arXiv preprint arXiv:1402.1892*, 2014.
45. J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 56–65, 2020.
46. R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

Appendices

Appendix A

For each individual, the following variables are noted:

- Gender
- Python experience
- Years of experience
- Education
- Internship
- Score
- Salary
- Offer history
- Location
- Recruitment status

A few rows of examples of the data is shown below.

Gender	Python experience	Years of experience	Education	Internship
Male	Yes	0	Graduate	No
Male	No	1	Graduate	No

Score	Salary	Offer history	Location	Recruitment status
5139	0	1	Urban	Y
4583	128	1	Rural	N

Appendix B

For each individual, the following variables below are noted. However, only the Gender, AISIyn, Shortlistedyn and Interviewed variables are completely filled in. The other columns lack information. OfferNY has 55 filled in entries, where AcceptNY and JoinYN has only 28 filled in entries.

- Gender : Assigned a code of 1 for Male and 2 for Female
- ATSIyn : Assigned 1 = Yes if candidate is an Aboriginal or Torres Strait Islander. Assigned 2 = No if candidate is a general applicant
- Shortlistedyn : Assigned 0 if rejected and 1 if shortlisted
- Interviewed : Assigned 0 if not interviewed and 1 if interviewed
- FemaleONpanel : Assigned 1 for Male only panel and 2 if a female member was present on the panel
- OfferNY : Assigned 1 if offer was made to candidate and 0 if not offered
- AcceptNY : Assigned 1 if accepted and 0 if declined
- JoinYN : Assigned 1 if joined and 0 if not joined

A few rows of examples of the data is shown below.

Gender	ATSIyn	Shortlistedyn	Interviewed
1	1	1	1
0	1	1	1

FemaleONpanel	OfferNY	AcceptNY	JoinYN
1	0	0	0
2	1	1	0

Appendix C

For each individual, the following variables below are noted.

- Application receipt date
- Applicant city PIN
- Applicant Gender
- Applicant Birth date
- Applicant Marital status
- Applicant Occupation
- Applicant Qualification
- Manager DOJ
- Manager Joining designation
- Manager Current designation
- Manager Grade
- Manager Status
- Manager Gender
- Manager DoB
- Manager Number of applications
- Business Sourced

A few rows of examples of the data is shown below.

Application receipt date	Applicant city PIN	Applicant Gender	Applicant Birth date
4/16/2007	844120	M	12/19/1971
4/16/2007	844111	M	2/17/1983

Applicant Marital status	Applicant Occupation	Applicant Qualification	Manager DOJ
M	Others	Graduate	11/10/2005
S	Others	Class XII	11/10/2005

Manager Joining designation	Manager Current designation	Manager Grade	Manager Status
Level 1	Level 2	3	Confirmation
Level 1	Level 2	3	Confirmation

Manager Gender	Manager DoB	Manager Number of applications	Business Sourced
M	2/17/1978	2	0
M	2/17/1978	2	1

Appendix D

This section states all the survey questions asked to the sub group recruiters. The overview of which fairness definition is covered in which question can be found in the table below.

Fairness measure	Question number
Individual fairness	18
False positive rate balance	21
False negative rate balance	21
Equal treatment	21
Predictive parity	22
Unawareness	9, 10
Counterfactual fairness	19
Overall accuracy equality	13
Statistical parity	11, 15, 16
Equal opportunity	20
Equalized odds	20

1. How many years have you been working within the field of recruitment?
2. For how many vacancies do you recruit each month (on average)?
3. How many candidates do you select for each vacancy (on average)?
4. Does your current company/organisation strive to have more women or men for the vacancies that you are recruiting for?
   - Yes, more women
   - Yes, more men
   - No, my company/organisations does not have such a policy
   - Does not apply
5. In which sectors(s) are you currently working? (for example: Healthcare, ICT, Education)

6. Have you been working with a software system that helps you with finding the best candidates to recruit by looking at their CVs?
7. Can you describe your recruitment process whenever you are using a software system that helps you find the best candidates?
8. Suppose that you are receiving 100 applications per week and you can only review 20 per week. There is software that can pre-select the best CVs. Would you use this software?
   – Scale 1-5
9. Suppose that you would use software to select the best candidates from CVs, how would you use it?
   – Let the software select and invite the candidates without checking the selected candidates
   – Let the software select the candidates, but check the selected candidates before inviting
   – Let the software show recommendations of candidates and make your own decision upon which candidates to recruit
10. Do you take into account a candidate's gender whenever selecting candidates?
    – Scale 1-5
11. Suppose that you would use software to select the best candidates from CVs and this software does not look at the candidates' gender (even though you do have this information). Do you think this software will discriminate based on gender?
    – Scale 1-5
12. Suppose that for a certain job application, 80% of the candidates are men. How many men/women would you select for interviews? You can invite a maximum of 10 candidates.
    men and 5 women
    men and 4 women
    men and 3 women
    men and 2 women
13. Suppose that an external organisation has tested the candidate selection software and says it is proven to be 99% accurate. Would you test the software yourself before implementing it into your company/organisation?
    – Yes
    – No
    – How would you test the candidate selection software before implementation within your company/organisation?
    – Why would you not test the candidate selection software before implementation within your company/organisation?
14. Suppose that an external organisation has tested the candidate selection software and says it is proven to be 90% accurate for female applicants and 95% for male applicants. Would you use this software?
    – Scale 1-5
15. Suppose that your company has tested the candidate selection software and says it is proven to be 99% accurate. Would you test the software yourself before implementing it into your company?

– Yes
– No
– How would you test the candidate selection software before implementation within your company/organisation?
– Why would you not test the candidate selection software before implementation within your company/organisation?

16. Suppose that you are using the software to find the best candidates. One day, it selects 9 men and only one woman for interviews. What would you do?
– Invited the 9 men and 1 woman
– Invite the 9 men and 1 woman but also additonal women
– Ask the software company to retune the software to make a more equal selection
– Disregard the software completely
17. Suppose that a vacancy states that it has a strong preference towards a specific gender and the software that you are using to find the best candidates is also incorporating this preference. Would you use this software?
– Scale 1-5
18. Would you compare your own candidate selection to the selection of the software that is trying to find the best candidates?
– Scale 1-5
19. Suppose that you would have two exact same candidates based on their CV, except that one is a man and the other is a woman. They are both fit for the job but the software only selects the woman. What would you do?
– Invite the man
– Invite the woman
– Invite both
20. Suppose that the candidate selection software has selected 10 candidates. The same software is used again, however the gender of the candidates is flipped (male switched to female and female switched to male). It then outputs the exact same 10 candidates. Do you think this makes the software fair?
– Scale 1-5
21. Suppose that you are using the candidate selection software, you invite 5 men and 5 women for each vacancy and after 10 filled vacancies you notice that 8 men filled the roles and 2 women. Do you think this is a fair outcome?
– Scale 1-5
22. Suppose that the candidate selection software selects 3 men with a bad CV and does not select 3 men with a good CV. For the same position, the software selects 1 woman with a bad CV and does not select 2 women with a good CV. Would you use this software?
– Scale 1-5
23. Suppose that you have to select candidates for a software development role. On which of the following variables with respect to best candidate selection would you base your choice?
– Python experience

- Work experience in years
- Internship
- Previous salary
- Offer history (if the candidate ever got a job offered at your company)
- Gender
- Graduated or not graduated
- Residence

Appendix E

This section states all the survey questions asked to the sub group potential applicants. The overview of which fairness definition is covered in which question can be found in the table below.

Fairness measure	Question number
Individual fairness	16
False positive rate balance	17, 18
False negative rate balance	17, 18
Equal treatment	17, 18
Predictive parity	19
Unawareness	11, 12
Counterfactual fairness	15
Overall accuracy equality	13
Statistical parity	8, 14
Equal opportunity	17
Equalized odds	17

1. How many years of working experience do you already have?
2. How many job applications did you participate in in the past 2 years?
3. What kind of jobs have you been applying for? (For example: IT consultant, HR manager, nurse)
4. What is your gender?
   - Female
   - Male
   - Prefer not to tell
   - Other
5. Have you ever been selected (as far as you know) by a software system that helps recruiters with finding the best candidates to recruit by looking at their CVs?
   - Yes
   - No
   - I don’t know

6. Suppose that a company offers you a choice between candidate selection by a human recruiter or a candidate selection software, which would you prefer?
   - Human recruiter
   - Algorithm
   - No preference
7. Overall, do you think selection of the best candidates will be less fair if companies use AI to support candidate selection?
   - Scale 1-5
8. Suppose that for a certain job application, 80% of the candidates are men. How many men/women should a company/organisation select for interviews? They can invite a maximum of 10 candidates.
   men and 5 women
   men and 4 women
   men and 3 women
   men and 2 women
9. Would you like to know if the software that the company is using to select the best candidates is treating candidates fairly based on gender?
   - Scale 1-5
10. Would you participate in an application process whenever software is used to select the best candidates based on their CVs?
    - Scale 1-5
11. Do you want to share your gender whenever you are starting an application process?
    - Scale 1-5
12. Suppose that you would share your gender information to the company that you are applying for and the company is using software to select the best candidates from CVs. Do you think this software will discriminate based on gender?
    - Scale 1-5
13. Suppose that an external organisation has tested the company's candidate * selection software and says it is proven to be 90% accurate for applicants of your gender and 95% applicants of the opposite gender. Would you want to participate in the application process?
    - Scale 1-5
14. Suppose that the company is using the candidate selection software to find the * best candidates. For your application process it selects 8 candidates of your opposite gender and one candidate of your gender, would you continue the application process?
    - Scale 1-5
15. Suppose that for the application process, the company switches all applicants' gender. So, male is switched to female and female is switched to male. With this information and the rest of the CV information, the software picks the best candidates. Do you think this would change your chance for being recruited?
    - Scale 1-5

16. Suppose that there is another applicant for the role that you are applying for and all the CV information the candidate selection software is looking at are the exact same for the both of you, except that the other applicant is from the opposite gender. You are selected, but the other person is not. Do you think this is fair?
    - Scale 1-5
17. Suppose that two male and a female applicant are selected by the candidate selection software, where one male has a bad CV and the others a good CV. In the pool of candidates there were 2 other males with a good CV and 2 with a bad one. Moreover, there was one other female with a good CV and one with a bad one. Do you think the selection is fair?
    - Scale 1-5
18. Suppose that the candidate selection software selects 3 applicants with the opposite gender as you and a bad CV and does not select 3 with the opposite gender and a good CV. For the same position, the software selects 1 applicant with the same gender and a bad CV and does not select 2 with the same gender and a good CV. Would you want to participate if this software is also used for your application selection process?
    - Scale 1-5
19. Suppose that you are applying for a software development role. Which of the following information on your CV do you think are important aspects to select the best candidate?
    - Python experience
    - Work experience in years
    - Internship
    - Previous salary
    - Offer history (if the candidate ever got a job offered at your company)
    - Gender
    - Graduated or not graduated
    - Residence

Appendix F

This section states all the survey questions asked to the sub group AI experts. The overview of which fairness definition is covered in which question can be found in the table below.

1. How many years have you been working within the field of AI?
2. Which kind of AI are you familiar with? (For example: NLP, Predictive models, Image classification)
3. Have you ever created a recruitment predictive model?
    - Yes (continue to question 4)
    - No (continue to question 5)
4. What did the process of creating your recruitment predictive model look like?

Fairness measure	Question number
Individual fairness	18
False positive rate balance	17
False negative rate balance	17
Equal treatment	17
Predictive parity	19
Unawareness	12, 13
Counterfactual fairness	15
Overall accuracy equality	14
Statistical parity	5, 9, 11
Equal opportunity	16
Equalized odds	16

5. Do you take fairness into account when creating AIs? (When not applicable, do not fill in this question.)
   – Scale 1-5
6. Suppose that for a certain job application, 80% of the candidates are men. How many men/women would you let an algorithm select for interviews? A maximum of 10 candidates can be invited.
   men and 5 women
   men and 4 women
   men and 3 women
   men and 2 women
7. Suppose that a recruiter is receiving 100 applications per week and can only review 20 per week. You are able to make an algorithm that would select the best candidates based on their CVs. Would you make this software?
   – Scale 1-5
8. How would you test the candidate selection algorithm?
9. Would you test your algorithm differently for each company that wants to use your algorithm?
   – Yes (go to question 10)
   – No (go to question 11)
10. How would you test your algorithm differently for each company?
11. How would you test your algorithm to be able to cover different companies?
12. Suppose that you created an algorithm to find the best candidates. One day, it * selects 9 men and only one woman for interviews. What would you do?
    – Leave the algorithm as it is
    – Check where the imbalance comes from
    – Check where the imbalance comes from and adjust the algorithm such that it becomes balanced
    – Create a completely new algorithm that does not discriminate based on gender
13. Suppose that you would create an algorithm to select the best candidates from * CVs, how would you recommend the use of it to recruiters?
    – Let the software select and invite the candidates without human oversight

– Let the software select the candidates, but check the results before inviting
– Let the software show recommendations of candidates and make an own decision on which candidates to recruit

14. Suppose that a company asks you to create an algorithm that will select males 60% of the time and females 40%. Would you cooperate?
– Scale 1-5
15. Do you think that a candidate selection algorithm needs to take into account a candidate's gender?
– Scale 1-5
16. Do you think that whenever a candidate selection algorithm does not take a candidate's gender into account, it cannot discriminate based on gender?
– Scale 1-5
17. Suppose that you are testing your candidate selection algorithm, it says that it is 90% accurate for female applicants and 95% for male applicants. Would you sell this software to companies?
– Scale 1-5
18. Suppose that your algorithm has the same candidate selection output whenever the gender of the applicants would be switched from male to female and from female to male. Do you think this makes the algorithm fair?
– Scale 1-5
19. Suppose that a company is using your algorithm to find the best candidates, they invite 5 men and 5 women for each vacancy and after 10 filled vacancies they notice that 8 men filled the roles and 2 women, and inform you about this. Would you change your candidate selection algorithm?
– Scale 1-5
20. Suppose that your candidate selection algorithm selects 3 men with a bad CV and does not select 3 men with a good CV. For the same position, the software selects 1 woman with a bad CV and does not select 2 women with a good CV. Would change your candidate selection algorithm?
– Scale 1-5
21. Suppose that there are 2 applicants with the exact same CV variables, except that one is female and the other is male. Only the male candidate is selected by your candidate selection algorithm. What would you do?
– Leave the algorithm as it is
– Check where the imbalance comes from
– Check where the imbalance comes from and adjust the algorithm such that it becomes balanced
– Create a completely new algorithm that does not discriminate based on gender
22. Suppose that you are creating a candidate selection algorithm for a software development role. On which of the following variables with respect to best candidate selection would you base your choice?
– Python experience
– Work experience in years
– Internship
– Previous salary

- Offer history (if the candidate ever got a job offered at your company)
- Gender
- Graduated or not graduated
- Residence

Appendix G

This section states the contributing code of this use-case project.

Model MLP_ET

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.20, random_state=42)
parameters = {'activation': 'logistic',
              'alpha': 0.0001,
              'hidden_layer_sizes': (15, 10),
              'learning_rate': 'constant',
              'max_iter': 1000000,
              'solver': 'adam'}
mlp = MLPClassifier(**parameters, random_state=0)
mlp.fit(X_train, y_train)


for i in range (0,101):
    th = i*0.01 #threshold
    df_test['MLP'] = np.where(df_test['MLP']
                             > th, 1, 0)
    fairnessscores,fairnessscores_male,
                   fairnessscores_female = model_run(df_test)
```

Model MLP_DT

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.20, random_state=42)
parameters = {'activation': 'logistic',
              'alpha': 0.0001,
              'hidden_layer_sizes': (15, 10),
              'learning_rate': 'constant',
              'max_iter': 1000000,
              'solver': 'adam'}
mlp = MLPClassifier(**parameters, random_state=0)
mlp.fit(X_train, y_train)


for i in range (0,101):
    th = i*0.01 #threshold
    th_m = 0.5
    th_fm = th
    df_test['MLP'] = np.where(df_test['Male'] == 1,
                              np.where(df_test['MLP']
                              > th_m, 1, 0),
                              np.where(df_test['MLP']
                              > th_fm, 1, 0))
    fairnessscores,fairnessscores_male,
                   fairnessscores_female = model_run(df_test)
```

Fairness value function

```
def fairnessmeasure_graph(measure_m,measure_fm):
    if (measure_m==measure_fm):
        f_value = 1
    elif measure_m <= measure_fm:
        f_value = (measure_m)/measure_fm
    else:
        f_value = (measure_fm)/measure_m
    return f_value
```