A study on bias against women in recruitment algorithms

Surveying the fairness literature in the search for a solution

Johan van den Berg[1]
Supervisors: J. Yang[1]**, S.E. Carter**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Johan van den Berg
Final project course: CSE3000 Research Project
Thesis committee: J. Yang, S.E. Carter, S.N.R. Buijsman, M.M. Specht

Abstract

Algorithms have a more prominent presence than ever in the domain of recruitment. Many different tasks ranging from finding candidates to scanning resumes are handled more and more by algorithms and less by humans. Automating these tasks has led to bias being exhibited towards different unprivileged groups, among these women. This has prompted a need to find solutions to this bias in order to achieve fairness in algorithms for everyone. This survey analyses the state of the literature on fairness and bias against women with a focus on recruitment algorithms. It has been found that a plethora of methods to achieve and measure fairness exist, with many of the technical methods having only been tested in a controlled environment and not in a production environment. Companies are not very forthcoming in sharing how they ensure fairness, which complicates the development of the field. There are many limitations to the current methods, however due to widespread usage of algorithms in the recruitment process, there is still a need for solutions to exist while these critiques are being addressed. It is vital for fairness to not only consider technical solutions, but also social solutions and to be aware of the limitations of each approach. Solving the issue of bias against women in algorithms requires sometimes realising that the best solution does not involve an algorithm at all, or in the case where an algorithm is applied, a critical engineer that is aware of the limitations and possibilities of different methods to achieve fairness. Future work is in addressing some major critiques of the current fairness literature and reducing the bias in society that often leads to algorithmic bias.

1 Introduction

Algorithms have become ubiquitous in the present day. It is almost impossible to imagine many tasks without them. Whether it is picking a movie to watch in the evening, scrolling through your timeline, or applying for a job, there will likely be an algorithm at some point in the process influencing the result. The usage of algorithms has strong advantages, but also similarly strong disadvantages. The benefits are clear, time and money can be saved by automating decision-making, risks are minimised, productivity is enhanced and human biases can be diminished [20]. It all sounds very promising, until the downsides are considered. The biggest being the possibility that the algorithm does not actually reduce bias, but rather reinforce and amplify it [8]. Bias can be exhibited by algorithms towards many groups, and is also often exhibited towards women [12]. This bias can also be present in the human resources field during recruitment [20], something which this survey looks into.

There are many examples of bias being exhibited against women because of algorithm usage in the recruitment process. The most prominent and often mentioned example is Amazon's recruiting tool. The tool was scrapped after it was found to show bias against women [7]. It was found that for software development jobs women were not being recommended by the system. This was because the system was trained on resumes from previous applicants to assess those of current job applicants. Since the field was largely male-dominated during the period from which the training data was gathered, the algorithm had learned to prefer males. Even though the company claims that the recommendations were merely a support tool for recruiters and they were not solely relied on, the problem persists. The usage of a gender-biased algorithm has led to discrimination of women, as they were treated differently based on their sex. This bias is something which should be mitigated, which is where algorithmic fairness solutions come in.

Due to the ever increasing use of algorithms, there has also been a significant increase in algorithmic fairness solutions. Research has been conducted into different kinds of bias and many fairness methods and definitions have been proposed, which Mehrabi et al. have summarised in a 2021 review [24]. The state of the research into social causes of, consequences of and solutions to gender bias in algorithms has been surveyed by Hall and Ellis [12]. Köchling et al. [20] have approached algorithmic bias and fairness in recruitment algorithms from a business perspective to get an overview of the state of the field. Finally, in contrast to Köchling et al., Fabris et al. [8] have taken a multi-disciplinary approach to research fairness and bias in the recruitment process to inform all stakeholders involved. This research will investigate mitigating bias against women in recruitment algorithms. Currently no research combines the existing knowledge on bias mitigation techniques, bias against women and recruitment algorithms. This paper closes this knowledge gap and adds to the existing knowledge base and identify further research avenues. Addressing this gap is important to ensuring fair algorithmic hiring for women. The question of how bias against women in recruiting algorithms can be mitigated using a combination of fairness methods will be addressed.

This paper will be structured as follows. Firstly, the knowledge gap and research method will be discussed. Afterwards the recruitment process and problems from algorithm usage in it will be outlined. The next section will discuss different kinds of gender bias and the causes of gender bias. Chapter 5 will give an overview of the current bias measures and mitigation methods. Chapter 6 gives recommendations on how to achieve fairness for women in the algorithmic recruitment process and outlines why it is important to look at combinations of fairness methods. Finally, there will be a discussion, a section on responsible research and a conclusion, which will also discuss possible further research.

2 Research Method

This chapter will provide an overview of the current literature into fairness, gender bias, the recruitment process and the intersection of those. The knowledge gap will be identified and the research methodology will be outlined.

2.1 Existing work and knowledge gap

In this section the current literature on fairness, recruitment algorithms and gender bias is discussed. The knowledge gaps are identified and the contribution of this paper is highlighted. Currently, there are reviews on fairness methods, fairness in recruitment and gender bias in algorithms, but no research so far into the intersection of these areas.

The previously mentioned review by Köchling et al. [20] in 2020 gives a good overview of the state of the field at that time from a business perspective. It discusses the algorithmic fairness literature in the HR field. Köchling et al. found that algorithms are increasingly used in the HR field and there is a possibility of discrimination and unfairness if the algorithms are blindly trusted. They identify that the knowledge of downsides in the field is underdeveloped in comparison to the adoption. The review aimed to create awareness of potential biases, inform about potential dangers of algorithmic decision making and identify future research directions. They go on to discuss the types of algorithms used, reasons bias might occur and how to measure fairness. After the literature search the findings are presented. In this section the literature is divided into the categories recruitment, selection and development. In the recruitment process it was found that social media was increasingly used, where recommender systems try to serve relevant ads to the job seeker and recruiter. Here past hiring decisions are used for training, which can be a source of bias. Another possible source of bias identified is the ability to target job seekers based on certain attributes which might be protected. Finally, it was found that job advertisements were not delivered in a gender neutral way even if intended, due to women being a more costly demographic to serve ads to. They further found that in the selection process the usage of CV and resume screening, online interviews and algorithmic evaluations had increased. Again here bias in historical data was found to be a large pitfall. They discuss some alternative techniques proposed. In HR development the research was found to be underdeveloped, but the same pitfalls apply in this area. In general it was found that there is a lack of transparency in the decision making and usage and more research in the area is needed. They conclude by calling on organisations to consider perceived fairness of algorithms, try to avoid bias and let humans make the final decision.

A more recent review by Fabris et al. [8] in 2024 takes a multidisciplinary approach to discuss fairness and bias in algorithmic recruitment. They describe the stages of the algorithmic hiring process more in depth. Institutional bias and technology blindspots are discussed, which accentuates the more interdisciplinary approach compared to Köchling et al. In contrast to Köchling et al., Fabris et al. also evaluate the literature on fairness measures and mitigation strategies. This paper focuses on technical solutions, highlighting future research avenues and supporting conceptualised understanding of algorithm usage in the recruitment process.

The 2023 review by Hall and Ellis [12] discusses societal causes and consequences of gender bias in algorithmic systems. They argue that most papers focus on technical causes, but the problems are often also societal, as such solutions should be socio-technical. The most common social consequence found was amplification of existing bias. Social solutions found were increasing diversity in design teams, increasing transparency, increasing awareness, using human-in-the-loop and implementing ethics into the design process. They note that their study is the first focusing on gender bias in the socio-technical framework. In the end they call on interdisciplinary collaboration in future work.

The paper by Mehrabi et al. [24] is one of the best-known papers in the field of algorithmic fairness. They conducted a comprehensive literature review on fairness in algorithms in 2021. In contrast to the previously discussed paper by Hall and Ellis Mehrabi et al focus on the technical aspect of algorithmic bias and fairness methods. In the survey they investigate real-world systems that have shown bias and go on to discuss the kinds of biases affecting algorithmic systems. Data to algorithms bias, algorithm to user and data to algorithm bias is discussed and the feedback loop phenomenon is described. This is when bias is fed into a system through the data, given to the user by the algorithm and then fed back into the system in the form of new data, as such perpetuating or amplifying the bias. They go on to discuss fairness definitions and methods, which will be discussed in chapter 5. In the end they express the need for further research into fairness methods outside of classification tasks and call on the research community to synthesise a single definition of fairness.

Mujtaba and Mahapatra in a 2019 survey discuss the types of bias occurring in the recruitment process and methods to mitigate this. In their literature review they found causes of bias in recruitment algorithms to include bias from the training data, label definitions, feature selection, proxy attributes and masking. They go on to discuss mitigation methods, of which a more complete overview is given by the paper by Mehrabi et al. [24].

The aim of this paper is to combine the existing knowledge in the literature mentioned above and further literature, to obtain an overview of gender bias in recruitment algorithms and the current state of the literature on mitigation methods. This angle currently remains unexplored and this research aims to fill this knowledge gap. It is important to address this gap as there exist real-world examples where bias has been exhibited against women in recruitment algorithms and it is important to ensure fair hiring for everyone, including women. After a review of the literature and analysing the state of the field, conclusions will be drawn and future research avenues will be identified to encourage further research on this topic.

2.2 Research method

As outlined above there is a need for work into avoiding gender bias arising from the usage of algorithms in the recruitment process. This survey focuses specifically on bias against women and aims to bring together the existing literature on this topic, fairness and the recruitment process. The focus on the recruitment process was chosen because of the availability of literature on this topic [20]. Due to limitations of social and technical methods in general and individually it is important to look at

combining methods. Conclusions will be drawn from the analysis and future research areas identified. The main research question will be as follows:

- How can bias against women in recruiting algorithms be mitigated using a combination of fairness methods?

To help answer this question the following sub questions have been formulated, with the chapter detailing the findings between brackets:

- SQ1: What problems arise from the usage of recruiting algorithms? (Chapter 3)
- SQ2: What kind of bias is exhibited against women in algorithms? (Chapter 4)
- SQ3: What kind of fairness metrics exist in the literature to assess bias in algorithms and what are the methods to mitigate bias? (Chapter 5)
- SQ4: What fairness methods could be applied to recruiting algorithms to achieve a higher degree of fairness for women? (Chapter 6)

The first sub question will be answered by giving an overview of the recruitment process and some examples of problems stemming from algorithm usage in this process, while the second will deal with types and causes of bias. The third question will be answered by giving an overview of the fairness literature and the final question by using the answers to the previous questions. These questions will be answered using a critical literature review. This method was chosen due to time constraints and because it is a good method to get an overview of the current state of the field and future research avenues. Limitations of this method are that it is hard to reproduce and that some works might be overlooked. To mitigate the risk of overlooking critical works, multiple databases were used in the literature search, including Google Scholar, the Leiden University Catalogue, IEEE XPlore and ArXiv. References from the literature gathered were also assessed for suitability. Literature found to be highly influential was analysed for relevance and included when relevant. After an outline of this paper was made, a search was conducted for literature relevant to each section. In the most cases, the most widely used literature was chosen. Additionally, as this is a high-level overview, the literature most relevant to this work is literature reviews.

The knowledge gap will be closed when the main research question is answered. The sub questions have been structured in such a way, that the main question can be answered using the answers to the sub questions. The sub questions can be considered answered when an overview of the relevant literature has been made and the relevant connections have been made to arrive at a conclusion.

3 Analysis of algorithm usage in the recruitment process

The following chapter will discuss the problems surrounding the usage of recruitment algorithms and gives an answer to SQ1. The first section will discuss where algorithms are used in the recruitment process. The second section will discuss problems stemming from this usage. Finally, some concluding remarks are given based on the findings.

3.1 Algorithm usage in recruitment

A survey in 2023 found that at least 97% of fortune 500 companies were using an applicant tracking system, while noting that the other companies might also be using tracking systems, but that this was not detectable [27]. These systems aim to optimise the hiring process by automating resume reviews and candidate searching. A notable figure in the usage of these systems is that over 70% of resumes are not reviewed by humans anymore [34]. The figures show a widespread use and adoption of algorithms in the recruitment process, as well as a reliance on them, as such it is important to assess possible bias introduced, perpetuated and exacerbated by the usage of these systems. There is a vast array of applications of algorithms in the process of finding a job or a candidate which are not described here as the goal is to give an understanding of the vast array of use cases in the field.

As mentioned in the introduction, algorithms can save time and money by automating parts of the recruitment process and they even have the possibility to reduce bias. Humans have the tendency to exhibit bias in the recruitment process, for example trough prejudices and personal beliefs, which is something possibly mitigated by algorithms if they are used in the right way [20]. In this way the algorithms can be a powerful support tool of HR employees to aid them in making objective and consistent decisions. Take the example where there is an opening for an IT job and there are two applicants, one unqualified male and one qualified female. If the recruiter then chooses the male, because in their mind they are better suited for the job (due to their being male), then the recruiter has exhibited bias. If an algorithm is made in such a way that it does not exhibit bias, then the algorithm could prevent the recruiter from making this biased choice, as the recruiter would have to substantiate the choice for the male over the female, while the algorithm has selected the female on the basis of qualifications. As such algorithms can be a powerful tool to promote fairness in hiring. The economic savings and the possibility to reduce bias are the main driving forces behind the adoption of such algorithms [26].

The recruitment process can be divided into sourcing and selection, finding candidates and choosing among the candidates respectively [8, 20]. Important to note is that the stages are fluid and as such algorithms for one stage can often be used in another. The literature on the recruitment process is most well-developed in the sourcing stage [8]. As this survey focuses on the recruitment process, algorithm usage in HR development will be outside the scope. A visual overview of the recruitment process can be found in figure 1.

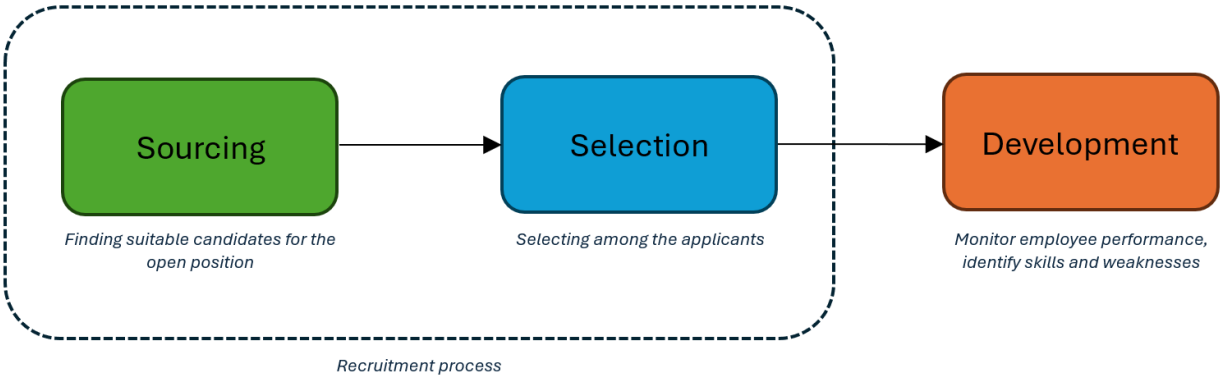

Figure 1: Overview of the stages of algorithmic hiring.

In HR recruitment the goal is to find a suitable candidate. From the recruitment side this can happen by posting a job and awaiting applications or actively looking for candidates to approach. A job seeker can also wait to be approached and make it known they are looking for a job or they can actively search for relevant job postings. In reality there is often a combination between an active and passive search [20]. The search for the candidate or job posting by the recruiter or the job seeker often happens online [34]. The advertising or searching is done mostly on employment platforms, such as Indeed, or social media, such as Facebook. Some platforms, such as LinkedIn, offer both social media and employment platform capabilities. This type of system uses algorithms to decide what postings and advertisements to show to candidates or what candidates to show to recruiters in searches and often also ranks the possibilities for the user [20]. The job advertisements posted might also make use of algorithms by using a language model to write them [8] and can be recommended to users by an algorithm based on certain attributes [22]. Candidates can be recommended to a recruiter based on their resume, which is screened automatically [29].

Selecting a candidate happens after the recruiting process. A list of candidates has been compiled and in this step one or more candidates from the list have to be selected. CV and resume screening can also be applied in this stage of the process to rank applicants and recommend them to the recruiter. To assess the candidates games, background checks, questionnaires, video interviews and chatbots can be used to extract information to subsequently analyse using an algorithm. The results can then be used to rank and exclude candidates based on skills, personality, preferences and suitability [8]. The skills and personality can also be used to evaluate the possible placement in the team as well as likelihood of acceptance of an offer.

3.2 Examples of problems with the usage of algorithms in the recruitment process

As said before, the usage of algorithms in a HR context can lead to discrimination and unfairness. There exist many examples of this, some of which are discussed below. The examples were taken from literature used for other sections of this paper. This section gives an overview of problems that have arisen or might arise from the usage of these algorithms and is meant to give an idea of the complications involved.

Firstly, consider the example of human bias in the previous section. The prevention of a biased choice hinges on the absence of bias in the algorithm. For multiple reasons making an unbiased algorithm is a difficult if not impossible task. First of all, as discussed in section 5.4, there exists no consensus on what is fair. An algorithm might be fair in one sense, but not in another. Secondly, take the example of the recruiters choice. If this happens often and these decisions are then encoded in a dataset, which is used to train an algorithm, the algorithm will mimic this behaviour. In this example historical bias is embedded in the data. Additionally, even if bias is not present in the data it could come from the design of the algorithm. This problem is elaborated on in chapter 4. The result of this is that if the necessary care is not taken in the algorithms development, bias could even be exacerbated compared to human decision making [34]. A good example of bias from historical data in practice is the Amazon example referred to earlier. Since in the past females were not hired for tech jobs, they were also not selected by the algorithm [38]. This shows that this example is not just theoretical, but has become reality.

A study by Lambrecht and Tucker in 2018 [22] found that ads in the STEM field were less likely to be shown to women, even when the ad was meant to be gender neutral. They found this to be because serving women ads is more costly than showing them to men. This is because of cost-optimisation in the algorithm. This real-world case of bias against women is a good example of a problem with algorithm usage in the search for candidates.

On Facebook companies are allowed to select sensitive attributes of users to target them for advertisement [1]. Some examples of sensitive attributes are gender, ethnicity and age. These attributes are often by law forbidden to discriminate on and in combination with other factors are as such considered sensitive or protected [4]. Kim and Scott [18] discuss a case where a job-seeker did not see the ads for a relevant job, as he was outside of the age category targeted by the ad. They note that excluding someone because of their age is illegal in the United States. This targeting in job advertisements means that in advance

a large portion of possible applicants is excluded [18]. Even choosing seemingly neutral attributes of users to target them on might lead to bias, since these might act as proxies for the sensitive attributes. Proxy variables are non-sensitive attributes have a correlation with the sensitive attribute [28].

A final problem with the usage of algorithms is a possible lack of personal contact when solely relying on the algorithm, which is related to the perception of fairness in the decision making [20]. Even if an algorithm can be deemed fair, a participant in the process could still have the perception it is not. The result is then that this individual is left with a feeling of unfairness after the process. Perceived fairness also has an effect on the willingness to accept a job offer even if selected [20]. It was found that regardless of the actual outcome of the decision, human, or algorithm-human decision making is more often perceived fair than pure algorithmic decision making [23]. It is therefore crucial to not solely rely on the algorithm, but also involve humans in the process.

The above are some examples of problems with recruitment algorithms to give an idea as to the scope and size of the problem. From these issues it is clear that there are concerns with the usage of algorithms in recruitment. However, as stated above there are clear advantages to the usage of such algorithms and in practice they are already widely adopted. This means that there is a need for attention to making these solutions fair for everyone.

3.3 Conclusion

The aim of this chapter was to give an overview of the recruitment process and problems arising from algorithm usage in it. The examples above clearly demonstrate some of the issues that arise. The main reasons algorithms are often deployed in the recruitment process are that they can save time and money and might reduce bias. From the examples it is clear that even though they have the possibility to reduce bias, they can still exhibit it. This bias has the potential to be even worse than the bias exhibited by a human in the recruitment process. It is thus vital to evaluate whether an algorithm is actually needed and whether it is actually fair. The next section will outline the kinds of bias that can be exhibited by these algorithms.

4 Bias against women in algorithms

Before being able to address bias it is important to understand that it can have many different causes and can exist in many different forms. This chapter aims to create an understanding of bias, it does so by first discussing different types of bias and then discussing the causes of this bias. Some takeaways are given in the conclusion. The chapter gives an answer to SQ2.

4.1 Types of bias

Bias can take many different forms, as such it is important to understand the kinds of bias, their causes, and consequences before being able to mitigate bias in algorithms. The survey by Mehrabi et al. [24] will be used throughout this section because of their clear overview of different forms of bias. Bias stemming from the data and the algorithm design process will be discussed. An overview of the types discussed can be found in appendix A.

Firstly types of bias arising in the data will be explained. When the data is fed into the algorithm, it will lead the algorithm to reflect this bias, leading to discrimination or unfairness [20]. A major cause for bias in datasets is related to the collection of data. The way of measuring and sampling a population can lead to bias [24], this is known as measurement, representation or sampling bias. This bias occurs when one group is over- or underrepresented in the dataset due to over- or under-sampling in the data collection, which can lead to problems in the learning of the algorithm. Take for example a dataset of past hiring decisions with mostly men in it. This might cause lower accuracy for women, leading to unfairness. Bias can also occur in the data when attributes that are relevant to the decision making are not included in the model, this is called omitted variable bias [24]. An example could be when trying to train a classifier for predicting salary, that does not include education level. Bias stemming from collection practices is known as statistical bias [25]. Important to note is that data can still contain bias even if the collection is done properly and the data perfectly represents the target population of the system, this is due to societal bias [25]. This type of bias is often called historical bias [24]. When for example women systematically receive lower wages, then even if the data reflects the reality, there is still bias in the data in the form that women get paid less. Training an algorithm on this data means that this bias will be perpetuated. It is important to understand these biases in order to find ways to address them.

Bias can also arise from choices made in the design process of the algorithm. Due to this an algorithm can exhibit bias even if there is no bias present in the input data. If the bias is added purely by the algorithm, this is called algorithmic bias [24]. This can occur for example when using a biased estimator in the model. Another type of bias stemming from the design process has to do with the difference between the target population and the user population. It might be that over time the population or the societal values have changed or that the statistics from the target population do not resemble those of the actual user population, this is then called emergent or population bias respectively [24]. An example would be when a model is trained on hiring decisions from a tech company and that same model is then used in hiring decisions at a production plant. There will then be a large difference between the population it was trained on and actually used on. In the design process of an algorithm it is important to be aware of these possible biases and to keep evaluating the algorithm during its lifespan.

The types of bias outlined above are the most relevant ones found and therefore outlined in this survey. A more elaborate overview of types of bias can be found in the survey by Mehrabi et al. [24].

4.2 Causes of gender bias

Aside from taking many different forms, bias can also have many different causes. Some were already outlined in the previous section through examples. This section presents an overview of causes of gender bias. Good to note is that even though this section focuses on causes of gender bias, the causes mentioned are largely universal.

Firstly, it is important to mention discrimination, as it is a source of bias and simultaneously an effect of bias. A distinction has to be made between explainable discrimination, when differences in outcomes between groups can be explained and justified by attributes, and unexplainable discrimination, when there is unjustified discrimination towards a certain group [24]. The last case is often illegal. An example of explainable discrimination is when males make more than females, but the females work less hours, thereby explaining and justifying the discrimination. Unexplainable discrimination can be split into direct and indirect discrimination [24]. Direct discrimination means that a non-favourable decision is based on a protected variable. An example is when a recruiter has to choose between a similarly qualified man and woman and chooses the man solely based on gender. When an algorithm is trained on data containing these decisions it will reflect this discrimination, the dataset then contains (historical) bias. Some attributes of an individual cannot be used in decision making, often by law. These are called sensitive or protected variables, this also includes gender. Indirect discrimination means the decisions are seemingly neutral and not based on the protected attribute, but the protected groups are still discriminated against, often because of proxy variables. This is when the protected attribute has a correlation with other variables [26]. An example of this type of bias in-practice is amazon's hiring algorithm, which even with gender removed, discriminated against women by assigning applicants from all-women's colleges lower scores [7]. The educational institution served as a proxy for the protected variable here.

Technical causes of gender bias often arise from the training data, label definitions, feature selection or proxy attributes [26]. Say for example that a recruiter labels a previous applicant a 'good' hire, there is then no information on what factors attributed to this label 'good', which could cause bias when this label is used in the decision making. If features which are not relevant to the outcome in the real world are included in the model, bias might arise when these features have an impact on the outcome of the model. Problems with the dataset are a central theme in the causes of bias. This bias can be both technical and social. The previous section on types of bias also clearly outlines some problems that can arise in the dataset.

Social causes of gender bias in algorithms are mostly related to the design of algorithms and the datasets used, as found in a review by Hall and Ellis [12]. In the design process most often a lack of diversity in the development team and lack of awareness were found to be the cause of gender bias. These problems are intertwined with each other, as a homogeneous team will have similar knowledge and thereby similar weaknesses. The awareness within the team of certain issues could then be lacking, which means a lack of diversity contributes to a lack of awareness. Bias can also be institutional, when the institution's practices, structure, customs and norms have a negative effect on the disadvantaged group [8]. This institutional bias can as outlined in the previous section also be present in the dataset. Another problem with the dataset is a lack of diversity, this was found by Hall and Ellis to be the most commonly identified social cause for gender bias in datasets.

Lastly, important to note is the feedback loop phenomenon. This is when a biased outcome from the model caused by bias in the data or the algorithm is then used to improve the model. The biased outcome is then included in the dataset and fed to the model, thereby perpetuating the bias [24]. This means this phenomenon is a cause of bias in itself.

The above causes of bias are the most common causes, some other causes which are not repeated here are mentioned in the previous section on types bias. An overview of the causes discussed can be found in appendix B. A take-away from this section is that bias does not have one cause, but rather a variety of causes. In a dataset this means that bias will often not be from one source but from multiple sources, something which should be taken into account.

4.3 Conclusion

This chapter has presented an overview of the kinds of biases against women that can arise from algorithm usage. As said, the bias present in society is reflected by data. Usage of this data then causes the algorithm to reflect this bias, perpetuating the discrimination of women. Even then if the data perfectly reflects society, it can still be biased, something which is important to be aware of when developing algorithms. Aside from bias present in society, bias can also be introduced in the data collection and algorithm design process. This all means that bias can take many different shapes and has many different causes. It is therefore important for design teams to try to address any bias to make algorithms as fair as possible or even refrain from using an algorithm in certain cases to make sure fair decisions are made. The next chapter will discuss different methods to address and measure bias in systems.

5 An overview of the fairness literature

This chapter will provide an overview of the fairness literature and provide an answer to SQ3 on fairness measures and methods. The reviews by Fabris et al. [8], Mehrabi et al. [24] and Verma and Rubin [36] have been critical in shaping this chapter. First off, the different definitions of fairness will be discussed, followed by technical methods to improve fairness. Social methods to improve fairness and some critiques on the current fairness literature are then discussed and finally some conclusions are drawn.

5.1 Fairness definitions

This section will give an overview of the most common definitions (also measures, notions and metrics) of fairness used currently. In the context of philosophy and psychology people have already tried to define fairness long before the existence of computer science and the definition of fairness has been a much debated topic over the course of time [24]. For a history of fairness one can refer to the overview given by Hutchinson and Mitchell [15] in 2018.

There exist different categorisations of fairness definitions in the literature. Mehrabi et al. [24] group notions into group, subgroup and individual fairness notions. Fabris et al. [8] only consider group fairness measures, as they found there was a lack of research into individual and subgroup methods in the recruitment field. They classify the group fairness notions according to what kind of variables the measures take into account, they call these flavours, an idea from Mitchell et al. [25]. They have divided the measures on a high level into outcome, accuracy, impact, process and representational fairness, with outcome fairness encapsulating most terms. These groupings will be mentioned throughout this section. For an explanation on some of the terms used here one can refer to the paper by Verma and Rubin [36].

Below, the different definitions of fairness are outlined. These definitions of fairness are the most common and were taken from Mehrabi et al. [24], Verma and Rubin [36] and Fabris et al. [8].

- **Demographic parity**: This group fairness method is also known as statistical parity and has an extension often referred to as conditional statistical parity [36]. It is also related to disparate impact (indirect discrimination), which uses the ratio, while this method uses the difference [8]. Demographic parity is based on the predicted outcome alone and is considered to be the most simple fairness notion [36] and is widely used. It states that membership of a protected class should have no correlation with the decision by ensuring equal positive prediction rates across groups. A limitation is that this method does not always ensure fairness. This measure allows qualified applicants from one class, but also unqualified applicants from the other class to receive a positive outcome [13] for the purpose of having equal rates. This has as the result that this notion can impair the utility of the algorithm, especially if the target value to predict has a correlation with the sensitive attribute that is not discriminatory.
- **Equalised odds and equal opportunity**: these two methods were proposed by Hardt et al. [13] as an alternative to demographic parity, since they considered this method flawed as outlined above. Equalised odds entails that the true positive rates and false positive rates should be the same across different values of the protected attributes. Equal opportunity calls for true positive rates to be the same among the protected and unprotected groups. These group fairness methods are based on predicted and actual outcome [36] and as such fall into the outcome fairness class [8].
- **Accuracy equality**: According to this notion of fairness, a classifier can be deemed fair if the protected and unprotected group have equal probabilities of being assigned to the correct group. This group metric is based on the predicted and actual outcome [36].
- **Treatment equality**: To satisfy this notion of fairness, the ratio of false positives and false negatives should be the same for both the unprotected and the protected group. This group metric is again based on the predicted and actual outcome [36].
- **Test fairness**: This group metric is based on the predicted probabilities and the actual outcome and is satisfied if individuals in both the unprotected class and protected class have an equal probability of correctly belonging to the positive class [24]. This method is the combination of balance for the positive and negative class [31].
- **Mean absolute error**: This measure compares the group-wise accuracy by taking the average of the absolute error for each individual in the group [8].
- **Sensitive AUC**: There might be information about a sensitive variable stored in proxy attributes, this method is used to measure this relationship. It does so by training a classifier based on the non sensitive features to predict the sensitive variable and evaluating the accuracy of the classifier [8]. Fairness is achieved when the accuracy of the classifier is very low. This method is related to demographic parity in the sense that it is also based on the notion that membership of the protected class should not have an impact on the decision making.
- **Subgroup fairness**: The idea behind this definition of fairness is that it can capture the relational structure in a domain, taking into account social, organisational and other connections [24]. Subgroups are different combinations of protected attributes forming a (sub)group (for example higher educated women) within a bigger group (e.g. women). The aim is to combine group and individual fairness by picking a group fairness definition and seeing whether it holds for a collection of subgroups.
- **Fairness through (un)awareness**: In contrast to statistical measures outlined above, these notions take into account other attributes besides the protected one and are categorised as individual [24] or similarity based measures [36]. Fairness through unawareness states that a classifier is fair if no sensitive attributes are explicitly used in the decision making [36]. Fairness through awareness is the notion that similar individuals should have similar classification, so the distance between the outputs should be at most the distance between the individuals. This requires creating distance metrics, which could become a source of discrimination by itself [36].

Fairness metric	Mehrabi et al.	Verma and Rubin	Fabris et al.
Demographic parity	Group	Predicted outcome	Outcome
Equalised odds and equal opportunity	Group	Predicted and actual outcome	Outcome
Accuracy equality	Group	Predicted and actual outcome	Outcome
Treatment equality	Group	Predicted and actual outcome	Outcome
Test fairness	Group	Predicted probabilities and actual outcome	Outcome
Mean absolute error	Group	-	Accuracy
Sensitive AUC	Group	-	Process
Subgroup fairness	Subgroup	-	-
Fairness through (un)awareness	Individual	Similarity-based	-
Counterfactual fairness	Individual	Similarity-based	-

Table 1: Overview of the classifications of the fairness metrics per paper.

- **Counterfactual fairness**: This individual fairness method falls into the class of definitions based on causal reasoning [36]. These reasoning's use a graph with nodes representing attributes and edges representing relationships between them. A graph is deemed fair if there is no path from the protected attribute to the predicted outcome with only proxy attributes (attributes that can be used to predict the sensitive attribute) [36]. This is based on the intuition that a decision is fair if it is the same in both this world and the counterfactual world, so the world where an individual has exactly the same attributes but the protected attribute is flipped [24].

These methods can be used to measure fairness and extensions are possible to consider multiple protected variables instead of one [6]. An overview of the methods and the classifications of them can be found in table 1. As mentioned previously, Fabris et al. focus on group fairness measures, as they have not found other work in the HR field, as such this could be a fruitful avenue for further research. They also identified a lack of knowledge into representational fairness, measures in this class aim to quantify stereotypes. For example when in a job description there is bias embedded in the usage of some 'masculine' or 'feminine' words.

Something to note here is that the group fairness methods can be classified as oblivious, as stated by Mitchell et al. [25]. This means that they depend only on the observed data and as such are defined through only the features, outcomes, scores and decisions [25]. The individual fairness methods are classified as incorporating additional context, this provides a way to map social goals onto mathematical formalisms [25]. This idea is further elaborated by Fabris et al. [8] who defined multiple flavours to divide the measures into. They name flavour as one dimension and go on to define the dimensions conditionality (accepting group differences if they can be explained), granularity (finer granularity means measuring on more operating conditions), normativity (setting precise targets fro group-wise quantities), interpretability and multinary (accounting for more than one sensitive attribute). These are seen as desirable qualities for a system to have. Some fairness measures mentioned by Fabris et al. [8] are extensions of these methods into recommender systems. The base measures are outlined here to give an intuitive insight into the measures used and an overview of the field without going too much into depth.

Important to note is that some definitions of fairness are incompatible with each other and can thus not be satisfied simultaneously [24]. Therefore, careful consideration is needed as to which fairness measure to use in a system, a problem which depends highly on the use case. Ruf and Detyniecki [31] have made an attempt to address this problem by making a decision tree to aid in the selection of a definition to use. This tree and problems with it will be discussed in chapter 6. Finally, as can be derived from this section, reviews take very different approaches into grouping fairness methods and deciding which metrics to include. This highlights the difficulty of creating a single definition of fairness [24].

5.2 Fairness improvements using technical methods

The algorithms in the literature can be divided into three classes: pre-, in- and post-processing. The classifications are based on when in the process the algorithm is applied, this can happen before, during or after the model training respectively [8]. The three classes are elaborated on below.

- **Pre-processing**: Methods in this class take place before training the model and can only be used if the mitigation algorithm is allowed to modify the training data [24]. Methods in this class aim to remove bias from the data by removing non-relevant protected or proxy attributes and/or modifying features leading to bias [26].
- **In-processing**: This is the class of methods applied during the training of the model. A substantial increase of fairness can be achieved when algorithms are applied in this stage [26]. However, these methods can only be applied when one is allowed to change the objective function or impose a constraint [24].
- **Post-processing**: When there is no access to the training data or training process of the model, algorithms can be applied after training. This class of algorithms reassigns the outcomes based on some function [24] or provides transparency in the decision making process by providing counterfactuals [26]. Besides assisting the employer, the counterfactuals can also be provided to the applicant to aid them in improving themselves.

Several toolkits exist that implement fairness metrics and methods. A prominent example often mentioned in the literature is IBM's AI Fairness 360 toolkit [3]. This offers a variety of pre-, in- and post-processing methods as well as different fairness metrics. Using this versatile toolkit, different methods and metrics can easily be compared against each other. Another popular toolkit is Lime [30], which was found by Mujtaba and Mahapatra [26] to be the most popular fairness toolkit on Github. This toolkit is aimed at trying to understand the decision making process in black-box machine learning models. The features with the most influence on the model decisions are identified by the toolkit to attain this goal [30].

In the literature there is often a focus on a single protected attribute [6] in the improvement methods. In practice however, an individual can be a part of multiple protected groups simultaneously and optimising for only a single protected attribute can cause problems. Consider the findings by Chen et al. [6], they optimise fairness for one attribute and then measure the effect on another. Take the protected attributes sex and race. When they optimised for sex and did not consider race, it was found that the subgroup non-white males received more unfavourable outcomes and the subgroup white females received more favourable outcomes. The optimisation has now caused more unfairness regarding race. Due to this possible amplification of bias when optimising for only one protected attribute, it is important to consider multiple protected attributes when applying algorithms to increase fairness.

Chen et al. [6] have researched the current methods for intersectional fairness improvements in a survey from April 2024. Of the following methods, four were found by Chen et al. [6] to be the best at mitigating bias while considering multiple protected attributes. The other three were taken from Fabris et al. [8] and seem to be the most widely regarded in the recruitment field.

- **Reweighing**: This pre-processing method aims to remove discrimination from a dataset by assigning weights to the tuples in the training data [17]. No tuples have to be changed in this method. The idea is reducing discrimination while maintaining the class probability. The weights are calculated by dividing the expected probability to see an object with a certain sensitive attribute and class by the actual observed probability.
- **Rule-based Scraping/Substitution**: These pre-processing methods are aimed at proxy reduction, reducing the information in other attributes about the sensitive attributes [8]. In text data all words referring to the sensitive attribute are removed or changed. For example when considering gender, the word 'his' is removed or changed to 'theirs'.
- **Adversarial Inference**: The goal of this in-processing method, like the above method, is proxy reduction. It achieves this by modelling an adversary trying to predict the sensitive attribute from a representation of the individual derived from a neural network [8]. The goal is then to minimise the accuracy of the adversary.
- **MAAT**: The regular version of this method works by training two models, one optimising for fairness and one optimising for performance. The output of both models is a vector of the probabilities of the input object belonging to each class [5], these are then combined into one. The multi-attribute variant works by training a model for each protected attribute. This algorithm takes into account both performance and fairness and needs access to the training data.
- **Equalised Odds Processing**: This method optimises for equalised odds during the training of the model alongside accuracy. Since it does not need access to the training data, this method is the best alternative to reweighing, MAAT and FairMask [6]. The performance is very similar to these methods which do require access to the training data.
- **FairMask**: This method trains models to predict the sensitive attribute based on the other attributes. It then uses the outcome of these models to change the sensitive attributes in the test data [6].
- **DetGreedy**: LinkedIn developed and published this method for re-ranking search results. This method is deployed in practice at LinkedIn to ensure gender-representative ranking of qualified candidates [10]. The method re-ranks items from top to bottom using the most relevant candidates from underrepresented groups. In testing it was found to have show an increase in fairness and subsequently implemented. Unclear are the actual in practice consequences of this algorithm, but if we look at the example above where race was the unconsidered attribute, one can immediately see a possible issue. This publication is one of the only ones outlining the application of a fairness method on a large scale production system [2].

An overview of these selected methods is given in table 2. The table outlines the approaches, sources, family and measures used per method. A note on the measures used is that these were used in the proposal to measure the effectiveness of the proposed method. Different methods could also have been used in most cases, except for where the algorithm depends on a certain notion of fairness. This also shows clearly which methods are most popular.

What the methods reweighing, equalised odds processing, MAAT and FairMask have in common is that when they consider multiple attributes instead of one protected attribute, that the accuracy stays relatively stable, but the precision and recall are greatly affected [6]. For further explanation on precision, recall and accuracy see Chen et al. [6]. The other methods above have not been tested in such a fashion, but theoretically the same will apply to them. Chen et al. show that fairness comes at a premium, these methods attain fairness at the cost of more misclassifications.

Important to note is that Fabris et al. [8] found in their literature survey that the literature on fairness methods in the recruitment field is mostly focusing on proxy reduction techniques currently. It was found that these techniques do not offer fairness improvements when the bias in the dataset is small and that it is unclear whether these techniques actually offer benefits when applied in practice and not just on a test-set. Thus it is important to keep considering and researching other applicable methods in this field.

Fairness method	From	Family	Approach	Measures used in method proposal
Reweighing	[6, 17]	Pre	Differential weighting of training data	Demographic parity
Rule-based Scraping/ Substitution	[8]	Pre	Proxy reduction	sAUC
Adversarial Inference	[8]	In	Proxy reduction	sAUC
MAAT	[5, 6]	In	Combining individually trained models	Demographic parity, Equalised odds, Equal opportunity
Equalised odds processing	[6, 24]	Post	Adjust output labels	Equalised odds
DetGreedy	[8, 10]	Post	Output re-ranking	Demographic parity, Equalised odds
FairMask	[6]	Pre, Post	Adjust output labels	Equalised odds

Table 2: Overview of selected technical fairness methods.

5.3 Fairness improvements using social methods

Hall and Ellis [12] argue that as causes of gender bias in algorithms are socio-technical, solutions should also be socio-technical. This section will discuss the social solutions to bias in algorithms, in contrast to the technical solutions in the previous section.

As stated in the previous section, providing an applicant with counterfactuals can help them improve themselves and provide transparency. Along with this, providing applicants with clear feedback actually improves the perceived fairness of a system [26]. Giving the applicant feedback is largely a social solution and while it does not improve fairness directly, it does improve transparency and perceived fairness. Thereby improving the acceptance of the usage of technology in the hiring process. A class of social solutions looks at improving due process. It was found that improving fairness, accountability and transparency was the most important solution in this category [12]. Providing an applicant with counterfactuals is a good example of how this can be done.

Most literature on social solutions to bias names solutions aimed at improving the algorithmic design process [12]. Increasing diversity in the design team is among the most common methods. It is argued that having diversity in the design team can help prevent algorithms from exhibiting gender bias, something which is partly substantiated by the findings that gender diversity in software development teams leads to more creativity and better decision making [19]. This claim can be further substantiated by the intuition that by introducing more diverse perspectives, the influence of cognitive bias on the algorithm is reduced [16]. Another method to improve the design process is by increasing awareness of bias [12]. If one is not aware of the possibility of bias occurring, they will not address it. As such raising awareness of fairness in algorithmic design is an important part of bias mitigation. Human-in-the-loop is also a popular method to address fairness concerns [12]. In this method the algorithm is continuously audited by a human during the development and after deployment to ensure fairness. As mentioned earlier, perceived fairness is also an important part of achieving fairness in algorithmic hiring. The human-in-the-loop method partly implements this. Finally, the last notable method to improve the design process is integrating ethics into it.

The social methods mentioned are all closely related and intertwined. Increasing diversity in the development team will lead to increased awareness, as a more diverse team will have different blind spots [12]. More awareness means that there is knowledge of the possible issues arising from algorithm usage, which can lead to calls for a code of ethics, human-in-the-loop approach and increased transparency. Adopting a code of ethics could also lead to increased transparency and the adoption of human-in-the-loop, as the code of ethics will call for responsible application of algorithms. Using human-in-the-loop means biases can be identified and mitigated at the source, leading to increased awareness of issues. These connections are a good thing, as it means that after the first step improvements to fairness will continue.

Fairness method	Approach
Diversity in development team	Increasing the number of diverse developers or designers in the team.
Increase awareness	Make users and developers aware of the possibility of bias occurring.
Human-in-the-loop	Human experts evaluate the algorithm continually during the development and deployment and make adjustments when necessary.
Integrate ethics into design process	Adopt a code of ethics to adhere to in the algorithmic design process.
Improve transparency	Be transparent about algorithm usage and how the result was achieved.

Table 3: Overview of selected social fairness methods.

As will be further outlined in the next section 5.4, technical methods have their limitations. As such it is important to also consider the social solutions. This class of solutions looks at improvements outside of those to the algorithm itself. A concrete example of why it is necessary to look beyond technical methods is one from section 4.1. Take the example that women systematically receive lower wages. Even if a technical method is applied which is able to perfectly balance the data

or the algorithm, the societal structures that have led to the imbalance remain unaddressed and unidentified. Using for example human-in-the-loop, an expert can identify the problem and make the responsible management aware such that they can address the problem at the core. Social solutions are mostly limited by their complexity and cost. Changing the composition of the team to be more diverse involves hiring more employees, raising awareness requires time and funds, having a human audit the process requires a capable engineer to spend time doing so, integrating ethics involves creating a code of ethics and continually assessing whether it is adhered to and finally, improving transparency also involves spending time and continually finding areas to improve on.

In conclusion, bias cannot solely be mitigated using only technical methods. It is therefore important to also take into account the social structures leading to bias and the social solutions, some of which are outlined in this section.

5.4 Critiques on current fairness literature

As demonstrated in the previous sections, the fairness literature is not without its limitations. Besides the already discussed limitations, there are some more instances where the literature falls short. Some critiques of the dominant view on fairness in algorithms are discussed in this section.

Green and Hu assert that the current fairness literature assumes that fairness is "constituted by satisfaction of the statistical constraints" [11, p. 2] or "reducible to constructing mathematical summaries of individuals' attributes that admit comparison between persons on a single standardised scale of similarity" [11, p. 2]. What they are saying is that the literature is preoccupied with satisfying mathematical formulations of fairness. This is a limited approach, something which Weinberg [37] also identified in a survey of fairness critiques. The main limitation of this approach is that it fails to take into account the broader social and moral context. As such, what may be fair according to the definitions of fairness might not be fair in the broader societal contexts. A limitation further underscored by the incompatibility of fairness metrics mentioned earlier, since the community labels this incompatibility as the impossibility of fairness. However, as Green an Hu argue, the community misdiagnoses the problem, as this merely shows that there are no easy answers, not that there are no right answers.

Further problems arise from the reliance on quantifiable features of society. When relying solely on the data and not taking into account the broader social context of the data, the non-quantifiable values will not be taken into account. An example is the gender-imbalance in household work, which is not measured. More work in the household is done by women, giving them a disadvantage on the job market [8]. Green and Hu [11] conclude that this reliance results in giving undue consideration to certain values, solely because they are quantifiable. They go on to note the issue with the reliance on historical data in machine learning. As bias and unfairness is present in society, this is also present in data from society. Even if society has improved in terms of fairness and bias, historical data will not reflect these changes, as they will have a positive effect in future data. An algorithm trained on this historical data will make decisions based on that, these decisions will then also not reflect the societal changes.

Relevant to both this section and the earlier section on fairness definitions is the COMPAS debate. COMPAS is an algorithm deployed in a criminal justice setting that assesses potential recidivism risk. In 2016 an analysis was published by ProPublica where they identified that the tool was racially biased as did not satisfy equal false positive rates by race [25]. The company replied by stating that the tool was fair as it did satisfy positive predictive rates. Further research assessed the tool against multiple measures of fairness with differing results. This debate shows the problem with the current definitions of fairness, what one sees as fair, the other does not. As such there is a need for a harmonised definition of fairness, however as Mitchell states: "there can be no harmony among definitions in a world where inequality and imperfect prediction are the reality" [25, p. 153]. In contrast to this Köchling et al. [20] mention synthesising a definition of fairness as an open question, once again symbolising the lack of agreement in the algorithmic fairness research.

As mentioned previously, an argument often driving the implementation of an algorithmic system in the hiring process is that of reduced bias through the usage of an algorithm. Köchling et al. [20] confirm this in their literature survey. As also discussed earlier, the algorithm does also has the potential to perpetuate and worsen the bias. The reason this is reiterated is that often companies do not provide information on how they ensure fairness in their systems, something which Bakalar et al. [2] from Facebook identified. It is very well possible that many companies are applying some sort of fairness method, but are simply not transparent about it. This lack of transparency also hinders the development of the field, as methods cannot be tested in deployed applications, but only in a controlled environment. Additionally, it means that knowledge on the real-world implications of these systems is difficult to come by. Companies sometimes even offer full-scale services to automate the hiring process, while the question whether or not algorithms can actually be fairer than humans remains unanswered [20]. For example, upon reviewing the website of the applicant tracking system Workday with almost a 40% market share [27], it was found that while they mention having diverse experts and using human-in-the-loop [35], it remains unclear how they train their systems and actually ensure fairness. They state this information is available to customers, it remains however unavailable for the applicants, who have to use their systems. This example also ties in with the perceived fairness of algorithmic hiring systems, as transparency in the usage and development of these systems can help improve the acceptance of algorithm usage in recruitment.

Lastly, a central issue in the machine learning fairness field identified by Weinberg [37] is technological solutionism. Fairness is achieved through technological solutions, which, as outlined above, might not even be fair in their social context. This issue goes further than just the fairness field and is actually more inherent to the algorithmic field. Algorithms are applied and

subsequently cause these fairness issues, which are then attempted to be solved using technical solutions. The root of the problem is then not the bias, but rather the application of the algorithm, which has caused the bias. This application has the potential to "crowd out other forms of knowledge and inquiry" [37, p. 82]. Selbst et al. phrase this very clearly when they state that there is a "failure to recognise the possibility that the best solution to a problem may not involve technology" [33, p. 63].

The issues outlined above are by no means an exhaustive list of critiques of the fairness literature. A clear overview of some critiques is given by Weinberg [37]. This review was also employed in shaping this section. The limitations discussed are important to note and it remains a future challenge to address some of the critiques in order to attain fair machine learning. As there are not yet any solutions to these issues, it is important to still look at the current methods as the wide usage of algorithms calls for action now. Which means there is a good argument for research into the critiqued methods.

5.5 Conclusion

As outlined in this section making algorithms fair is a difficult task without a silver bullet. Solutions can be social and technical or somewhere in the middle and for the best resolution a combination of these should be considered, something which is outlined in the next section. The plethora of methods to assess whether an algorithm is fair illustrates the difficulty of the task at hand, the same is true for the methods to address bias. Some critiques were outlined, where some very good points were brought up about limitations of the current approaches. As stated previously, it is very important to address these limitations, however, in the meantime other concrete solutions are needed. This is because in practice algorithms are widely applied in the recruitment process and to ensure fairness and prevent discrimination, action should be taken now.

6 Applying fairness methods to recruitment algorithms

This chapter will provide an answer to SQ4 on what fairness methods to apply to recruitment algorithms to achieve fairness for women and why combining fairness methods is important. The insights derived from the previous chapters are analysed and some final remarks are discussed in the conclusion.

6.1 The ideal solution

Sometimes the best way to mitigate bias is to not use an algorithm at all. Stating that the best solution to bias in algorithms is not using an algorithm might seem simplistic, it is however more nuanced than that. Before starting on the implementation of an algorithm it is important to evaluate whether it is actually necessary to implement and what the impact would be. Then if it is actually implemented it is important to continually assess the algorithm. While no definitive proof exists that algorithms actually exhibit less bias than humans [20], it would be the best solution to steer clear of them. Logically, bias against women in algorithms cannot exist when no algorithms are used. Then it all comes down to human bias. Algorithms of course have the potential to mitigate this bias, however they often perpetuate or exacerbate it [8]. As such this should be the best solution. The problem is however that the world is not perfect and in the current society cost is a driving factor behind many decisions [20]. As algorithms are often simply cheaper than humans, there is also a big push for the implementation of these algorithms to save cost. The result of this is that there is a need to assess and mitigate bias in these systems to ensure fair hiring for everyone.

6.2 The practical solution: how to achieve fairness for women in recruitment algorithms?

It is important to first consider the social solutions and then the technical solutions. This is because bias often stems from society [12], thus it logically follows it is better to address the societal structures that led to the bias rather than address the bias only. So directly fighting the disease not just the symptoms. The most important solution in general to bias in algorithms is to have a competent critical engineer to be embedded in the design process and to evaluate the system after deployment, known as human-in-the-loop. This method is related to integrating ethics in the design process, having diversity in development teams, raising awareness and increasing transparency. If a qualified engineer aware of the issues with algorithms is employed for the purpose, they can stand up for the rights of others and critically evaluate the system during the design process and after deployment. For example, if they find historical bias against a certain group in the data, then they can voice these concerns to the management, which can address the bias. In doing so, the societal structure leading to bias is addressed. The critical engineer will also call for including diverse people in the process, thereby achieving diversity even if the team is not. They can help increase awareness of possible pitfalls in algorithm design among the design team. For example, if a critical engineer was employed in Facebook's advertising team, then they might suggest that to raise awareness of bias in advertisements a solution could be to put information about discrimination when posting an advertisement about a job offer, or they might suggest it should not be possible to target users on certain features at all. Besides awareness of social solutions, the engineer should also be aware of the technical possibilities and have knowledge on when to implement what method. If they find the algorithm or the data to be biased, they can recommend and insist on methods for mitigation. The critical engineer is therefore key to mitigating bias in algorithms. The issue is that there is a shortage of these engineers, supported by the need to still raise awareness of fairness issues [12]. Another problem is that employing these engineers and implementing these solutions can be costly. Furthermore, the effects of the social solutions are in most cases not immediately reflected in the data as the methods are often a timely process. This means that these methods, while being very useful, do not provide the definitive solution. For further increases of fairness the technical methods can be used alongside the social methods. The two classes of methods have differing limitations, which makes the combination of them especially suitable for achieving fairness.

Before mitigating any bias, it needs to be measured. The flow chart by Ruf and Detyniecki [31] is a good conception of a way to select a measure to use. Even though the idea is good, there are some issues within this specific context. The problem is that in hiring ground-truth labels may not be available [21]. This means regardless of policy and base rates the result according to the flow chart is some parity-based measure. This family of measures coincides with the practice of affirmative action [9]. The issue being that even when the flow is followed without a diversity policy, the end result can be the implicit application of such a policy through the usage of a parity-based measure. Evidence points to success of affirmative action policies [32], however demographic parity should still be applied with care and while knowing the implications. The chart contains every group fairness method mentioned except for mean absolute error and sensitive AUC. The mean absolute error method also relies on ground truth [8], it is a good metric to use if the accuracy of the model is vital. If no ground truth is available, other methods may still be used, but one must be aware of the limitations [8]. Say one uses equalised odds on data with historical bias against women. If one assumes the decisions made to be correct (ground-truth) and optimises a predictor for this ground truth and for equal false negative and true positive rates across groups, then the resulting model will still contain the historical bias from the dataset. As such, measures based on ground truth are limited for data without reliable ground truth availability. If there is no desire to implement a diversity policy and no ground truth available, a measure such as sensitive AUC can be used, which is also extendable to allow multi-class fairness assessment [14]. This method is popular as it coincides with the legislation against direct discrimination [8] and differs from demographic parity which targets indirect discrimination. This class of methods related to proxy discrimination should still be further analysed however, as it unclear whether the benefits hold after deployment [8]. As individual and subgroup measures remain understudied in hiring [8], no recommendations about the application can be made.

If bias is found in the data, the data collection should first be looked at. Non-representative data is often a source of bias and improving the data collection practices is the solution most often implemented in practice [2]. Two methods not to apply are DetGreedy and Adversarial Inference. These two methods take into account only one protected variable, which as outlined earlier has the possibility to increase bias in subgroups. As such it is better to avoid these methods until a multi-attribute extension is published and assessed. If there is no access to the training data equalised odds processing can be used, as it is the only method among the remaining methods to not require access. Choosing among usage of reweighing, MAAT and FairMask can be done per use case by assessing the usefulness of the methods. As the methods can all be used with any fairness metric [5, 6, 17], the methods should be compared against each other using the chosen measure and the most appropriate one should be selected. After choosing a method to apply, it is important to keep assessing the system when it is in use and stop using it if there are fairness issues found. When dealing with text data (such as resumes) it is advisable to at least use rule-based scraping, as this aims to make the text gender neutral, this has a small impact [8], but can help in mitigating gender bias and requires no access to the training data or the model. This method can be used in conjunction with other methods [8].

In short, numerous methods exist to mitigate bias against women and there is not one silver bullet. An overview of methods discussed can be found in appendix C. A combination of methods can be applied to achieve the fairest solution for all. The key to fairness being responsible application of algorithm. This does not only refer to the application itself, but also encompasses the development process and the assessment after deployment. In every stage it is important to have a team that is able to critically evaluate the algorithm and data and is willing to ensure fairness for everyone.

6.3 Conclusion

The previous section shows that the most important solutions for mitigating bias against women in recruitment are continually evaluating whether an algorithm is desirable and making sure the process is monitored by humans. In business decisions cost plays a big role and automating decision making is often simply cheaper than employing humans to do the same task [8]. As such, when dealing with systems that will become reality regardless of desirability from a social point of view, the best solution found was to at least make use of the human-in-the-loop construct to mitigate possible bias. If a suitable auditor is employed, this also partly accounts for the other social methods discussed. The bias found by an auditor can be mitigated using different solutions based on the type of algorithm deployed. The most commonly applied method would be to improve the representation of subgroups in the data. Based on the use case technical methods can be considered if bias is found using a selected measure. In conclusion, there exist many methods to attain fairness and to apply them in a way such that bias against women can be mitigated takes critical engineers willing to stand up for the rights of others.

7 Discussion

Despite the exploded research into fairness methods, the implementations of them in production systems is lacking behind. The most often adopted solution to a biased system is to simply add more training data for the underrepresented class [2]. This clearly shows that there are large issues with the algorithmic fairness literature, as outlined by the critiques. The lacking adoption of the technical methods of achieving fairness shows that there is a need for other solutions, as was discussed earlier. The implication of this is that it is important for development teams to adopt social solutions to the problem.

While technical fairness methods often provide band-aid solutions to large societal problems. The social methods discussed aim to address some of these societal problems, however these methods often do not translate into right-now solutions. Take for example raising awareness, this is a timely process and while this process is taking place algorithms with fairness concerns

needing to be addressed are already widely used. This is where the technical solutions come in. This suggests that in development teams should be aware of the implications and try to address any possible fairness issues. However, despite the utility of these technical solutions, the question remains whether this is the way to achieve fairness. Despite the plethora of methods proposed, the main concerns about achieving some statistical notion of fairness remain unaddressed.

Currently, companies are not very forthcoming into revealing the ways in which they have adopted algorithms and the fairness approaches used, which means public knowledge on how fairness methods work in a real-world environment is limited. Transparency is a main theme in the algorithmic fairness literature, as a lack of transparency also hinders the development of the field. This research identified this problem and calls for increased transparency in the algorithmic hiring space.

This review did not cover all methods and measures in the algorithmic hiring field, which could be construed as a limitation. Additionally, due to the design of the paper only the literature deemed critical was reviewed. There is the possibility of having missed key research because of this. It is however improbable that research not included has come to very different conclusions. In the literature, and especially also in the most recent literature, the critiques of fairness remain mostly unaddressed, as such the conclusions from the review will hold. Finding a way to definitively achieve fairness for everyone therefore remains an open question.

A high-level view has been adopted in this review, without examining any measures or algorithms in practice. The limitation of this way of working is that the fairness methods discussed might compare differently to each other in a practical sense than a conceptual sense. It could for example very well be that DetGreedy also performs well in fairness among subgroups. Then it has been ruled out on a conceptual level, but performs well in practice. However, the performance of such methods would then have been evaluated in very specific circumstances. The advantage of the current methodology is that methods which might work well in certain circumstances can be ruled out based on conceptual limitations.

In the review it has been stated that sometimes no algorithm should be involved at all. While it remains unclear whether algorithms can be fairer than humans, this might hold. However, as algorithms also have the possibility of being able to be fairer than humans, it is possible that the solution to bias in the hiring space is actually applying an algorithm instead of human review. Currently, this conclusion is not yet supported by the literature, however it is important to mention the possibility and the implications to the solutions discussed in this review.

The literature used for this survey has been combined and conclusions have been drawn from the results. This has yielded the conclusion that both technical and social methods alone do not provide a solution to the problem. As most literature focuses on either of the two, with exceptions of course, it is important to provide an overview of the possibilities the different literature offers. It is therefore logical to consider the possible combinations of methods offered by literature that can be applied to achieve fairness for women in recruitment algorithms.

Spreading awareness of fairness in algorithmic hiring is one of the areas this survey hopes to make an impact. As it has been identified that there is not one single answer to the problem of bias in algorithmic hiring, it is important that developers and users of algorithms in the hiring space are aware of the limitations of these algorithms. Especially important is the insight that sometimes the best way to ensure fairness is not using an algorithm and that applying one can have a large impact. When applied, it is important to make the algorithm as fair as possible for everyone. To achieve this, the knowledge of different techniques to achieve fairness and possible fairness issues should be widely spread.

8 Conclusion and Future Work

This review answers the question of how bias against women in recruiting algorithms can be mitigated using a combination of fairness methods. It has been found that many methods exist and there is no one solution to address the issue. This means looking at combinations of methods is important for ensuring fairness. Responsible application of algorithms is key. Especially social methods are critical to attaining fairness when an algorithm is used, as they in some cases address bias at the source. This bias would otherwise have been reflected in the data and would have to be addressed by technical methods. Current fairness literature has its limitations, but it has the potential to at least partly address bias in algorithms against women if the right combination of methods is used.

From this review it can be concluded that one of the most important future work areas is that of addressing the limitations and critiques of the current fairness literature. Addressing some of these concerns will not only have a large effect on fairness in recruitment, but rather everywhere algorithms are used. As such addressing these issues is key to achieving a fairer world for everyone. To address fairness issues in the hiring space it is critical that the knowledge of limitations of algorithms and fairness becomes more widespread. Related to this is the question of how to increase the transparency surrounding algorithms. Research into the best ways to spread knowledge of algorithms and fairness should be conducted.

A crucial insight to remember is that the fairness methods all have their limitations. Before applying any fairness method it is important to be aware of these limitations and make an informed choice about the methods to apply. The implications of demographic parity are especially important to consider. Finally, the key takeaway is that responsible application of algorithms is crucial to achieving fairness. Before development of the system possible implications of the system should be reviewed and an informed choice should be made about whether or not to continue to development. In development it is important to have critical engineers who closely evaluate the system and propose changes if necessary. After a system is deployed it is important to keep assessing the system to continually ensure fairness for everyone.

9 Responsible Research

This review has looked at the implications of algorithm usage in the recruitment process. The ethical aspects of this have been discussed throughout the work. A key takeaway from the work is that algorithms should be applied responsibly to ensure fairness for everyone, this could go as far as refraining from using an algorithm for certain tasks as discussed. The conclusions were logically derived from the information gathered and are as such reproducible.

A Types of bias

Type	Classification	Description
Measurement bias	Statistical	Bias stemming from using or measuring features in a certain way [24].
Representation bias	Statistical	Not all subgroups are represented equally in the data [24].
Sampling bias	Statistical	Subgroups are not sampled equally [24].
Omitted variable bias	Statistical	Relevant attributes not included in the model [24].
Historical bias	Societal	The data contains bias not stemming from collection practices, but from societal structures [25].
Biased estimator	Algorithmic	The estimator used in the model is statistically biased [24].
Emergent bias	Algorithmic	The population or the societal values have changed since the development of the algorithm [24].
Population bias	Algorithmic	Statistics from the target population do not match those of the actual user population [24].

Table 4: Overview of types of bias discussed in section 4.1.

B Causes of bias

Cause	Related to	Description
Historical bias	Dataset	Previous injustices in the world have become embedded in the data [24].
Training data	Dataset	This cause refers to bias present in the dataset in general and encompasses other terms mentioned [26].
Label definitions	Dataset	Unclear labels in the data can lead to bias [26].
Feature selection	Dataset	Features not relevant in the decision making are included or features relevant to the decision making are not included in the dataset and used to train the model. This can lead to bias [26].
Proxy attributes	Dataset	Information about the sensitive attribute is stored in other attributes [26].
Institutional bias	Dataset	An institution's practices, structure, customs and norms have a negative effect on a disadvantaged group [8].
Lack of diversity	Dataset	The dataset is not representative of all groups [12].
Lack of diversity	Development team	Lacking diversity in the development team means homogeneous knowledge and thus the same blind spots or biases [12].
Lack of awareness	Development team	The development team is not aware of the importance of certain issues [12].
Feedback loop	Development process	A biased outcome from the model caused by bias in the data or the algorithm is used to improve the model. The biased outcome is then included in the dataset and fed to the new model, thereby perpetuating the bias [24].

Table 5: Overview of causes of bias discussed in section 4.2.

C Solutions to bias

Method	When to use?
No algorithm usage	Whenever there is a risk of the algorithm being unfair. It is important to assess whether the cost saving will weigh up to the possible societal consequences of the algorithm application.
Critical engineer	At least one team member should be able to take on this role for ensuring fairness in development. The knowledge of fair algorithms in the development team should be invested in as much as possible. Application of this method can be limited by availability of knowledge and costs associated with it.
Improving data collection	This should be the first step to improving fairness when there is issues found with an algorithm. It can address many different forms of bias without having to change the algorithm.
Equalised odds processing	Whenever there is no access to the training data.
Reweighing, MAAT, FairMask	Per use case by assessment of the usefulness of the methods. Choose the best out of the three or none upon assessment.
Rule-based scraping	When dealing with text data.

Table 6: Overview of solutions discussed in chapter 6.

References

[1] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to skewed outcomes. *CoRR*, abs/1904.02095, 2019.

[2] Chloe Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, and Jiejing Zhao. Fairness on the ground: Applying algorithmic fairness approaches to production systems, 2021.

[3] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

[4] Simon Caton and Christian Haas. Fairness in machine learning: A survey, 2020.

[5] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. Maat: a novel ensemble approach to addressing fairness and performance bugs for machine learning software. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2022, page 1122–1134, New York, NY, USA, 2022. Association for Computing Machinery.

[6] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairness improvement with multiple protected attributes: How far are we?, 2024.

[7] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/idUSKCN1MK0AG/, 10 2018. [Accessed 28-04-2024].

[8] Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. Fairness and bias in algorithmic hiring: a multidisciplinary survey, 2024.

[9] James R. Foulds and Shimei Pan. Are parity-based notions of ai fairness desirable? *IEEE Data Eng. Bull.*, 43:51–73, 2020.

[10] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search: Recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery: Data Mining*, KDD '19. ACM, July 2019.

[11] Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*, 2018.

[12] Paula Hall and Debbie Ellis. A systematic review of socio-technical gender bias in ai algorithms. *Online Information Review*, 47:1264–1279, 03 2023.

[13] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

[14] Léo Hemamou, Arthur Guillon, Jean-Claude Martin, and Chloé Clavel. Don't judge me by my face : An indirect adversarial approach to remove sensitive information from multimodal neural representation in asynchronous job video interviews, 2021.

[15] Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. ACM, January 2019.

[16] Kristin N Johnson. Automating the risk of bias. *George Washington Law Review*, 87(6):19–12, 2019.

[17] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, Oct 2012.

[18] Pauline T. Kim and Sharion Scott. Discrimination in online employment recruiting. *Saint Louis University law journal*, 63(1), Dec 2018.

[19] Karina Kohl and Rafael Prikladnicki. Benefits and difficulties of gender diversity on software development teams: A qualitative study. In *Proceedings of the XXXVI Brazilian Symposium on Software Engineering*, SBES '22, page 21–30, New York, NY, USA, 2022. Association for Computing Machinery.

[20] Alina Köchling and Marius Claus Wehner. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 13(3):795–848, November 2020.

[21] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 275–284, New York, NY, USA, 2017. Association for Computing Machinery.

[22] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7):2966–2981, 2019.

[23] Maude Lavanchy, Patrick Reichert, Jayanth Narayanan, and Krishna Savani. Applicants' fairness perceptions of algorithm-driven hiring procedures. *Journal of Business Ethics*, 188(1):125–150, Nov 2023.

[24] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), 7 2021.

[25] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(Volume 8, 2021):141–163, 2021.

[26] Dena F. Mujtaba and Nihar R. Mahapatra. Ethical considerations in ai-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7, 2019.

[27] Sydney Myers. 2023 Applicant Tracking System (ATS) usage report: Key shifts and strategies for job seekers. https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems/, 10 2023. [Accessed 28-05-2024].

[28] Dana Pessach and Erez Shmueli. *Algorithmic Fairness*, pages 867–886. Springer International Publishing, Cham, 2023.

[29] Alejandro Peña, Ignacio Serna, Aythami Morales, and Julian Fierrez. Bias in multimodal ai: Testbed for fair automatic recruitment, 2020.

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[31] Boris Ruf and Marcin Detyniecki. Towards the right kind of fairness in ai, 2021.

[32] Simone Schotte, Rachel M. Gisselquist, and Tharcisio Leone. Does affirmative action address ethnic inequality? Technical Report 14, UNU-WIDER, Helsinki, Finland, January 2023.

[33] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA, 2019. Association for Computing Machinery.

[34] Nicholas Tilmes. Disability, fairness, and algorithmic bias in ai recruitment. *Ethics and Information Technology*, 24(2):21, Apr 2022.

[35] Kelly Trindel. Workday's continued diligence to ethical ai and ml trust. https://blog.workday.com/en-us/2022/workdays-continued-diligence-ethical-ai-and-ml-trust.html, Dec 2022. [Accessed 05-06-2024].

[36] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7, 2018.

[37] Lindsay Weinberg. Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ml fairness approaches. *Journal of Artificial Intelligence Research*, 74:75–109, May 2022.

[38] Galit Wellner and Tiran Rothman. Feminist ai: Can we expect our ai systems to become feminist? *Philosophy & Technology*, 33(2):191–205, Jun 2020.