Perceived Algorithmic Fairness using Organizational Justice Theory

An Empirical Case Study on Algorithmic Hiring

Guusje Juijn
5848423
g.juijn@students.uu.nl

First Supervisor:
Dr. D.P. (Dong) Nguyen
Utrecht University

Daily Supervisor:
Ir. N. (Niya) Stoimenova
DEUS, Delft University of Technology

Second Examiner:
Dr. M.M.A. (Maartje) de Graaf
Utrecht University

A thesis presented for the Master's degree in
Artificial Intelligence

Department of Information and Computing Sciences
Utrecht University
The Netherlands
April 2023

Abstract

Growing concerns about the fairness of algorithmic decision-making systems have prompted a proliferation of mathematical formulations aimed at remedying algorithmic bias. Yet, integrating mathematical fairness alone into algorithms is insufficient to ensure their acceptance, trust, and support by humans. It is also essential to understand what humans perceive as fair. In this study, I therefore conduct an empirical user study into crowdworkers' algorithmic fairness perceptions, focusing on algorithmic hiring. I build on perspectives from organizational justice theory, which categorizes fairness into distributive, procedural, and interactional components. By grouping participants based on the type of information they receive about several hypothetical recruitment algorithms, I find that algorithmic fairness perceptions are higher when crowdworkers are provided not only with information about the algorithmic outcome but also about the decision-making process. Remarkably, this effect is even observed when the decision-making process can be considered unfair, when gender, a sensitive attribute, is used as a main feature. By showing realistic trade-offs between fairness criteria, I find a preference for equalizing false negatives over equalizing selection rates amongst groups. Moreover, I discover a negative effect of selection rate differences and false negative rate differences on fairness perceptions. These findings contribute to the literature on the connection between mathematical algorithmic fairness and perceived algorithmic fairness, and highlight the importance of considering multiple components of algorithmic fairness, rather than solely treating it as an outcome distribution problem. Importantly, this study highlights the potential benefits of leveraging organizational justice theory to enhance the evaluation of perceived algorithmic fairness.

Acknowledgements

I would like to thank all people who helped me with this thesis project throughout the past eight months. First of all, I especially want to thank Dong Nguyen. I am really grateful for all the time you took to have meetings with me, help me with my questions, and provide me with constructive feedback when I got stuck. I appreciate your trust in letting me perform a crowdsourced experiment, as well as helping me with writing my first academic paper. I moreover want to thank my supervisors from DEUS, Niya, and Joao, for giving me the opportunity to gain experience in a growing AI company and taking the time to have weekly meetings to provide me with interesting insights and ideas. Special thanks to my fellow interns at DEUS and my thesis group at the University: it really helped to be surrounded by people going through the same process. I want to thank Rosanna Nagtegaal and Maartje de Graaf, for providing me with interesting insights into how to perform a user experiment. I moreover want to thank Sieuwert van Otterloo, for giving me relevant insights into the data set I used for this thesis. Lastly, I want to thank my family, Sam, and my friends for always believing in me and keeping me motivated throughout this process.

Contents

List of Tables

List of Figures

Introduction

Artificial Intelligence systems are increasingly being used to inform and make important decisions about human lives across a wide range of high-impact domains, such as criminal law, medicine, finance, and employment. Human decision-makers are increasingly being assisted by algorithmic systems in making critical decisions, like whether somebody should receive a mortgage, which medical treatment a patient should receive, or whether a defendant should be granted parole [1]. For example, a study carried out in 2018 into US banking institutions showed that 91% of the largest banks used deep-learning algorithms to make financial decisions about their customers [2]. Moreover, by 2014, the market for algorithmic job screening systems was already approximated at $500 million per year, with an annual growth of 15% [3].

These decision-making algorithms could offer numerous promising advantages to society. By automating time-consuming, mundane tasks and accelerating decision-making processes, these systems often increase efficiency, accuracy, and productivity, leading to important financial benefits [4,5]. Furthermore, algorithms have the potential to make more neutral judgments than humans, since they are not affected by emotions or other surrounding noise inherent to human decisions [6]. Algorithmic decision-making could therefore lead to more consistent, rational, and fair outcomes than traditional decision-making [4].

However, over the last couple of years, a considerable amount of literature has emerged that offers contradictory findings to this aspiration: algorithmic decision-making is increasingly being associated with discriminatory or unfair outcomes. Cases like COMPAS, the criminal risk assessment algorithm which was accused of being racially biased against black defendants [7], Amazon's recruitment model, which turned out to be biased against female candidates [8], and the research by Buolamwani and Gebru [9] into facial recognition systems, showing significant misclassification of black women, are some of the most canonical examples reporting algorithmic unfairness.

To counter harmful events such as these, ensuring algorithmic fairness has recently become a major area of interest within the field of artificial intelligence. Diversity, fairness, and non-discrimination have become key requirements in the EU Ethical Guidelines for Trustworthy AI [10]. Algorithmic justice and fairness have been highlighted in over 60 ethical guidelines for artificial intelligence [11]. Academic conferences such as ACM's FAccT (Fairness, Accountability, and Transparency) have come up, promoting more research into algorithmic bias [12]. These developments have led to the design of a whole landscape of fairness criteria: statistical expressions to embed fairness into algorithms. Furthermore, researchers have developed various bias mitigation algorithms, open source libraries, and auditing toolkits to measure, visualize, and improve different algorithmic fairness aspects [13].

Nonetheless, there are still large gaps between fairness researchers and machine learning practitioners in industry [12]. It has been proved that it is impossible to mathematically satisfy all the proposed statistical fairness criteria at once, since they are mutually incompatible. Therefore, a universal consensus on how to ensure algorithmic fairness is lacking: a one-size-fits-all solution does simply not exist [14]. More knowledge about what criteria to use in what context is hence needed, which exemplifies the fact that algorithmic fairness should not only be perceived from a technical viewpoint. We need to understand what humans perceive as fair, to ensure that algorithmic decision-making systems are accepted, trusted, and supported by humans. Therefore, multiple studies have started to acknowledge

that fairness is not purely an algorithmic concept, but a human construct [1,12,15,16].

However, the literature on human perceptions of algorithmic fairness so far frequently offers mixed or inconsistent results [1,17]. For example, some studies find that perceptions of algorithmic fairness are influenced by sociodemographic factors like age and education level [18], while others find opposite results [19]. Some studies find that people perceive human decision-makers as more fair [20], while other studies find a preference for algorithmic decision-making systems [21]. While some studies find a preference for straightforward fairness criteria like demographic parity, which requires equal selection rates between different groups [22], others conclude that people perceive more complex criteria, such as equal false positive rates between groups, as fairer [23]. Moreover, some studies find that people perceive the use of certain input features in a model, such as gender, as unfair [24], whilst others do not draw this conclusion [25].

To create more coherency and clarity in testing perceptions of algorithmic fairness, multiple fairness researchers have started to draw insights from a branch of psychology termed *organizational justice* [5,23,26–29]. This research area is concerned with fairness perceptions of decisions made about employees in organizational settings [30]. It divides fairness perceptions into three several different, but correlated components: distributive fairness, procedural fairness, and interactional fairness [31]. While distributive fairness is concerned with the fairness of outcome distributions (e.g., the amount of money paid to female and male employees), procedural fairness is concerned with the fairness of a decision-making process (e.g., the role of gender in resume screening), and interactional fairness is concerned with the fairness of the information provided about a decision-making process (e.g., the explanations provided for a layoff decision).

In this thesis, I will adopt this categorization in the context of the perceived fairness of algorithmic decision-making, focusing on algorithmic hiring. By approaching this topic through the lens of organizational justice theory, I will systematically investigate different aspects of perceived algorithmic fairness.

1.1 Motivation and Research Questions

Most of the research into algorithmic fairness perceptions focuses merely on one of the above-described components of fairness (for example, on distributive fairness, by investigating people's perceptions of different algorithmic outcomes, or on interactional fairness, by investigating the effect of explanations about algorithmic decisions). However, in this work, I aim to investigate *the effect of integrating these components on algorithmic fairness perceptions*. Additionally, I will investigate the link between mathematical algorithmic fairness and human algorithmic fairness perceptions. I will do so by examining *whether participants have a preference for either demographic parity or equality of opportunity* and by examining *whether there is a relation between fairness perceptions and, respectively, algorithmic selection rate differences and false negative rate differences*.

I will focus on algorithmic hiring, a context that is easily comprehensible for a lay public. While this area has seen increased interest in the integration of AI-enabled software, it has also witnessed raising concerns about the potential of AI to perpetuate or exacerbate existing biases [32–34]. As a result, it is classified as a high-risk area in the proposed EU AI act [35]. Moreover, there is no universal agreement on how fairness should be formalized in algorithmic hiring: for instance, certain recruitment algorithms proactively aim to increase diversity when ranking job candidates, while others do not [36]. As research has demonstrated that fairness perceptions during a hiring process play a critical role in job satisfaction, performance, and the relationship between employers and employees, obtaining insights into the perceived fairness of algorithmic hiring is of particular importance [37].

Toward that end, I will conduct an experiment using the crowdsourcing platform Prolific Academic to examine crowdworkers' fairness perceptions of several hypothetical recruitment algorithms. I will

study the following three research questions:

RQ1: *How do human fairness perceptions of a recruitment algorithm differ when only given information about the distributive fairness of the algorithm, compared to when given information about both the procedural fairness and the distributive fairness of the algorithm?*

This research question will be answered by grouping the participants based on the type of information they receive about the recruitment algorithms, according to the fairness components described in organizational justice theory. By calculating the average scores that participants assign to the algorithms, and comparing these scores across groups, I will investigate whether there are significant differences across the groups.

RQ2: *How do human fairness perceptions of a recruitment algorithm differ depending on whether it adheres to demographic parity or equality of opportunity?*

I will specifically focus on these two fairness criteria, as multiple studies propose these criteria are suitable in the context of algorithmic hiring [5, 33, 36, 38]. By showing participants graphs that report the trade-offs between selection rate differences and false negative rate differences between two gender groups, I will investigate whether participants have a preference for either equality of opportunity (i.e., equal false negative rates between both groups) or demographic parity (i.e., equal selection rates between both groups). Moreover, I will qualitatively analyze the rationales behind participants' fairness ratings to find out whether these affirm the results of the quantitative analysis.

RQ3: *To what extent are the selection rate differences and false negative rate differences between groups of a recruitment prediction algorithm related to human fairness perceptions of it?*

This research question will investigate the relationship between mathematical algorithmic fairness and perceived algorithmic fairness. I will answer this question by calculating the selection rate differences and false negative rate differences of the various recruitment algorithms shown to participants and comparing these to the participants' average perceived fairness scores. Additionally, I will employ an ordinal regression analysis, to predict fairness perceptions from selection rate differences and false negative rate differences.

In Section 4, these procedures will be outlined in further detail.

1.2 Definitions

In order to combat bias and discrimination in machine learning, it is essential to define when an algorithmic decision is unfair. Throughout this research, I will adopt the following frequently cited definition by Mehrabi et al. (2018) to broadly describe **fairness** in the context of algorithmic decision-making:

"Fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits" [39]

I will define **bias** according to the definition of Friedman and Nissenbaum (1996):

"A computer system is biased when it systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others" [40]

In this thesis, I will elaborate on several **fairness criteria**:

"A fairness criterion is a quantification of unwanted bias in training data or models" [41]

Moreover, I will use the terms **privileged group**, **unprivileged group**, **protected attribute**, and **favorable label**:

"(Un)privileged groups are groups (often defined by one or more sensitive variables) that are disproportionately (less) more likely to be positively classified" [42]

"Protected attributes define the aspects of data which are socioculturally precarious for the application of machine learning. Common examples are gender, ethnicity, and age" [42]

"The favorable label is a label whose value corresponds to an outcome that provides an advantage to the recipient (such as receiving a loan, being hired for a job, not being arrested)" [41]

1.3 Internship at DEUS

I am writing this thesis as part of a graduation internship at DEUS. This humanity-centered Artificial Intelligence company, consisting of interdisciplinary teams of engineers, data scientists, strategists, and designers, is established in 2020 and is based across three locations in Amsterdam, Porto, and A Coruña. DEUS helps organizations and companies with starting new artificial intelligence initiatives, launching pilot projects, and scaling these initiatives. Besides, DEUS also creates its own IT and AI products [43].

This graduation internship takes place within one of these product development initiatives: the Reliable AI platform. In this project, DEUS is creating a tool that can enable companies to deploy and maintain reliable AI models, by monitoring different aspects of deployed machine learning models with not only a focus on model performance but also on business impact and human insights and feedback. In this project, I focus on the human aspect of this monitoring tool, with a particular emphasis on algorithmic fairness and human perceptions of it. By performing a literature study into the empirical research into this topic and conducting experiments on human fairness perceptions, I hope to provide DEUS with relevant insights that can be used in creating the Reliable AI platform.

1.4 Thesis Outline

This thesis consists of seven sections. Section 2 introduces the theoretical underpinnings of algorithmic fairness. Section 3 presents an overview of related research on algorithmic fairness perceptions. Section 4 outlines the methodological framework of this study, including the details of the data (Section 4.1), model development procedures (Section 4.2), and empirical user study (Section 4.3). Section 5 describes the quantitative and qualitative findings of this study. In Section 6, these findings, as well as the study's limitations (Section 6.2) and directions for future research (Section 6.3), are further discussed. Finally, Section 7 presents the final conclusions of the thesis.

Theoretical Background on Algorithmic Fairness

This section will start by giving a succinct exploration of the different sources of algorithmic biases (Section 2.1) and the different mathematical criteria for algorithmic fairness (Section 2.2). It will proceed by explaining why it is impossible to satisfy all these criteria together at once (Section 2.3), and by providing a short overview of the technical approaches to mitigate algorithmic unfairness (Section 2.4). Altogether, this section will provide a structured theoretical background on algorithmic fairness, which will help in further examining the different viewpoints on this topic.

2.1 Algorithmic Bias

A first step towards preventing discriminatory and unfair algorithmic outcomes is to gain insight into how algorithmic biases emerge and where they come from. Multiple researchers have proposed different categorizations or frameworks to differentiate between the several forms of algorithmic bias. Mehrabi et al. [39], for example, define three different categories of algorithmic bias in the machine learning pipeline: bias in the data, bias in the algorithm, and bias emergent from the user interacting with the system.

First of all, when biased data is fed into a machine learning model, this might result in biased outcomes. Data can be biased due to different reasons. For example, when collecting data from a population, specific subgroups may be missed or under-represented, resulting in a sample lacking diversity and therefore not being representative of the population. An illustrative example of this phenomenon named *representation bias* is the demographic bias present in the large ImageNet data set used for classifying images: although this data set is intended for universal use, for example, only 1% of its images are taken in China, resulting in a lower classification performance of persons or objects in these images [44]. A special type of biased data is *historical bias*. This is one of the most complicated types of bias because it can emerge even after a perfect data sampling process [40]. It arises when the real world, as it used to be, or currently is, leads to biased models [44]. For example, when a data set contains only a little fraction of data points representing a certain minority group, this may result in a model that works less well for this group. In this case, even though the data reflects the reality accurately, we can question whether we want the model to reflect this reality [39]. Amazon's recruitment algorithm is an illustration of an algorithm that demonstrated historical bias. This algorithm was trained to evaluate job candidates based on resumes submitted over a period of 10 years. Since most of these resumes came from male candidates, as the tech industry was largely male-dominated, Amazon's algorithm learned itself to favor male candidates, by penalizing resumes containing the word 'women's' [8]. Because of historical bias, rooted human biases can be perpetuated: for example, in the case of word embeddings used for natural language processing algorithms, it has been demonstrated that profession words like 'doctor' or 'housekeeper' show a high association, or semantic relationship, with words representing respectively men or women. This can result in the reinforcement of human stereotypes, leading to discriminatory outcomes [45].

Bias can also be added by the algorithm itself, without formerly being present in the input data. In that case, the bias is caused by the inner workings of the algorithm, by specific design choices, or by data processing steps, such as feature selection or feature engineering. For example, when a machine

learning engineer tries to reduce the complexity of a model through hyperparameter tuning, by setting higher regularisation parameters, this generalizes the model but may end up eliminating subtle or rare patterns in the data. As a consequence, hyperparameter tuning can lead to a biased model that inadequately represents certain sub-populations. This is a form of *technical bias* [12].

Lastly, the third category of bias, *emergent bias*, or *user interaction bias*, emerges after deployment of the machine learning model [40]. This form of bias usually arises due to shifts in societal knowledge, demographics, or values over time. For example, this type of bias may arise when the population utilizing the model is different from the population presumed as users during the design phase. When an algorithm, for example, includes a large set of written instructions, a population lacking literacy skills may not be able to utilize it effectively, causing bias against nonliterate individuals [39,40].

2.2 Mathematical criteria for fairness in algorithmic decision-making

The fact that algorithmic fairness is a profoundly complex and many-faceted concept is reflected by the large landscape of definitions that try to grasp its meaning: with over 21 established mathematical formulas for fairness in binary classification problems, researchers have not yet come to a universal consensus on how to mathematically define what it means for a decision to be fair [46]. These mathematical formulas, commonly called fairness criteria, are broadly dividable into two classes: similarity-based, or individual fairness criteria, and statistical, or group-level fairness criteria [14]. In this section, I will discuss the most prominent criteria within these two classes. The terms *protected attribute* and *sensitive attribute* will be used interchangeably, referring to demographic features on which discrimination is not allowed, such as age, race, or gender.

2.2.1 Similarity-based fairness criteria

Similarity-based fairness focuses on the idea that individuals that are similar should receive similar predictions [14]. The most straightforward type of this class of fairness, and therefore the most frequently applied fairness criterion in organizations, is a blindness approach termed *fairness through unawareness*.

Fairness through unawareness: a classifier satisfies fairness through unawareness if sensitive attributes are left out from the data, and are therefore not used in comparing individuals [47].

In theory, this approach may sound promising. One could, for example, argue that by removing racial attributes from the data set, it would be impossible to discriminate on race. However, in practice, fairness through unawareness does not prevent discrimination. In fact, this approach may even contribute to the perpetuation of biases over time [5]. The reason for this is that many features have a tendency to be correlated with the sensitive attribute: these features serve as a *proxy*. For example, the feature 'zip code' may be a proxy for the sensitive attribute 'race' and the feature 'occupation' may be a proxy for the sensitive attribute 'gender' [42]. Therefore, even when a sensitive attribute is removed from the data set, a classifier will likely find one or more features that are *redundantly encoded*, or highly correlated, with this sensitive attribute—ending up with a classifier that essentially learns the same patterns [48].

Dwork et al. [49] argue that algorithmic fairness, rather, can best be described by treating individuals that are similar in a similar way, by actually being *aware* of their sensitive attributes:

Fairness through awareness: a classifier satisfies fairness through awareness if individuals that are similar with respect to a specific task receive similar predictions [49].

In this criterion, the concept of similarity is described by a similarity metric, that measures the distance between two individuals. This similarity is then compared to the distributions of their decision outputs. In this approach, however, the important choice of which similarity metric to use is difficult

and requires situation-specific expert knowledge. Another difficulty with this criterion is how to test it in practice: to test similarity between all individuals, an enormous search space would be required [14]. Taken together, these concerns make fairness through awareness a difficult measure to operationalize in practice.

2.2.2 Statistical fairness criteria

Rather than comparing persons at an individual level, statistical fairness criteria focus on treating persons that belong to a *protected group* (defined by a sensitive attribute) the same as persons that belong to any other group. To capture the different formulas belonging to this class, Boracas et al. [48] propose a taxonomy of statistical non-discrimination criteria consisting of three categories: *independence*, *separation*, and *sufficiency*. If we represent the sensitive attribute as S, the predicted outcome (the decision) as $\hat{Y}$, and the (true) outcome as Y, these three categories can be represented as follows:

$$Independence = \hat{Y} \perp S$$

$$Separation = \hat{Y} \perp S|Y$$

$$Sufficiency = Y \perp S|\hat{Y}$$

Throughout the following section, the running example of a recruitment prediction machine learning model, that hires job candidates, will be used to illustrate the different fairness criteria belonging to these three categories. In this example, the sensitive attribute S is the sex of the candidate (*female* or *male*), the predicted outcome $\hat{Y}$ is the decision of whether the candidate will be hired for the job ($\hat{Y} = 1$) or not ($\hat{Y} = 0$), and Y is the true outcome of whether the candidate is qualified for the job ($Y = 1$) or not ($Y = 0$). In this model, the *favorable outcome* is the desirable decision outcome: being hired for the job. In this specific example, females are considered to belong to the protected group.

Independence Starting with *independence*, it can be seen that the true outcome Y is not considered in the equation. So, a classifier satisfies independence, when the predicted outcome is statistically independent of the sensitive attribute, regardless of the actual outcome. Three fairness criteria falling under the independence category are *demographic parity* (a.k.a. *statistical parity*), *disparate impact* and *conditional statistical parity* [14,48].

Demographic parity: a classifier satisfies demographic parity when the percentage of favorable outcomes is equal for both the protected and unprotected groups [39].

Our example recruitment prediction machine learning model would hence satisfy demographic parity when female candidates have an equal probability of being hired for the job as male candidates:

$$P(\hat{Y} = 1|S = female) = P(\hat{Y} = 1|S = male)$$

Table 1 shows an example of a recruitment algorithm that satisfies demographic parity.

A related variant of demographic parity, that also considers the percentage of favorable outcomes, is *disparate impact*. Instead of requiring these percentages to be equal, disparate impact considers the ratio of favorable outcomes between the protected and unprotected groups [42]:

$$\frac{P(\hat{Y} = 1|S = female)}{P(\hat{Y} = 1|S = male)}$$

US law refers to this criterion with the "80% rule": according to the US Equal Employment Opportunity Commission (EEOC) guidelines, it is stated that if the selection rate of the protected group is less than 80% than that of the unprotected group, there is discrimination based on a protected attribute. In fairness literature, demographic parity and disparate impact are often described as useful fairness criteria in situations where it is desirable to enforce equality between two groups, for example, in the

case of historical biases in the data set [46]. However, one may argue it is sometimes more reasonable to also involve other factors in the decision-making process. This can be done by extending the definition of demographic parity into *conditional statistical parity* [50].

Conditional statistical parity: a classifier satisfies conditional statistical parity when the percentage of favorable outcomes is equal for both the protected and unprotected groups, controlled for a legitimate factor L [50].

In our recruiting example, this means that the model would satisfy conditional statistical parity when female candidates have an equal probability of being hired for the job as male candidates, given that they have, for example, the same skills or work experience, denoted by L:

$$P(\hat{Y} = 1|S = female, L = l) = P(\hat{Y} = 1|S = male, L = l)$$

Table 1: Example of a recruitment algorithm adhering to demographic parity. This algorithm adheres to demographic parity as the percentage of hired candidates is equal for both men and women (p = 0.5).

Candidate	**Predicted outcome $\hat{Y}$**
Woman 1	hired
Woman 2	hired
Woman 3	not hired
Woman 4	not hired
Man 1	hired
Man 2	hired
Man 3	not hired
Man 4	not hired

Separation Instead of conditioning on a factor L, it is also possible to condition over the true outcome Y. If the predicted outcome is conditionally independent of the sensitive attribute, given the true outcome, a classifier satisfies *separation* [48]. For separation to hold, hence, the true outcome should be available. A scenario in which the objective true outcome, or *ground truth*, can be determined is, for instance, a medical setting where algorithmic predictions are made about whether an individual has a specific illness that can be detected through a blood test. Our algorithmic recruitment prediction example, however, is an instance in which this objective ground truth is not available but is inferred, as the qualification of a job candidate is likely a subjective decision [3]. When adopting fairness criteria falling under separation, hence, it is important to properly define what is meant by the true outcome Y, and to use data that is as objective and trustworthy as possible [46].

The three most prominent definitions falling under this category are *predictive equality, equality of opportunity*, and *equalized odds* [46, 51, 52].

Predictive equality: a classifier satisfies predictive equality when the false positive rate is equal for both the protected and the unprotected groups [51].

In our running example, the recruitment prediction model would satisfy predictive equality if, for both female and male candidates, the probability that an applicant does get hired for the job but is not actually qualified for the job (a *false positive*), is equal. Mathematically, this also implies that the true negative rate is equal for both the protected and unprotected groups: meaning that the recruitment model would correctly reject unqualified female and male candidates (*true negatives*) at the same rate

as well:

$$P(\hat{Y} = 1|Y = 0, S = female) = P(\hat{Y} = 1|Y = 0, S = male)$$
$$P(\hat{Y} = 0|Y = 0, S = female) = P(\hat{Y} = 0|Y = 0, S = male)$$

Table 2 shows an example of a recruitment algorithm that satisfies predictive equality.

Table 2: Example of a recruitment algorithm adhering to predictive equality. This algorithm adheres to predictive equality as the probability of being hired, given not being qualified for the job, is equal for both men and women (p = 0.5). Note that this algorithm does not adhere to demographic parity, as the percentage of hired female candidates is 0.75 and the percentage of hired male candidates is 0.5.

Candidte	**Predicted outcome $\hat{Y}$**	**True outcome Y**
Woman 1	hired	qualified
Woman 2	hired	qualified
Woman 3	hired	not qualified
Woman 4	not hired	not qualified
Man 1	hired	not qualified
Man 2	hired	not qualified
Man 3	not hired	not qualified
Man 4	not hired	not qualified

The second notion of separation, *equality of opportunity*, closely resembles the former criterion but focuses on the false negative rate, rather than on the false positive rate:

Equality of opportunity: a classifier satisfies equality of opportunity when the false negative rate is equal for both the protected and the unprotected groups [52].

This would imply that the recruitment prediction model would incorrectly reject qualified female and male candidates (*false negatives*) at the same rate. Mathematically, this also implies that the true positive rate, or *recall*, is also equal for both the protected and unprotected groups: meaning that the recruitment model would correctly hire qualified female and male candidates (*true positives*) at the same rate as well:

$$P(\hat{Y} = 0|Y = 1, S = female) = P(\hat{Y} = 0|Y = 1, S = male)$$
$$P(\hat{Y} = 1|Y = 1, S = female) = P(\hat{Y} = 1|Y = 1, S = male)$$

Table 3 shows an example of a recruitment algorithm that satisfies equality of opportunity.

When determining the suitability of either predictive equality or equality of opportunity in a given situation, it is important to identify whether fairness between groups is more sensitive to false negatives or false positives. To illustrate, the criterion of predictive equality does not take false negatives into account. Hence, in a situation in which equal false negatives between groups are crucial for fairness, predictive equality is not a suitable criterion. For example, in algorithmic hiring, it would be unfair to disproportionally reject qualified candidates (false negatives). Therefore, predictive equality would not be a suitable criterion in algorithmic hiring, as it does not ensure equal false negatives. On the other hand, the criterion of equality of opportunity does not take false positives into account. Hence, in a situation in which equal false positives between groups are crucial for fairness, equality of opportunity is not a suitable criterion. An example of such a situation is algorithmic firing: here, it would be unfair to disproportionally fire qualified employees (false positives). Hence, in algorithmic firing, equality of opportunity would not be a suitable criterion, as it does not ensure equal false positives [3].

Table 3: Example of a recruitment algorithm adhering to equality of opportunity. This algorithm adheres to equality of opportunity as the probability of not being hired, given being qualified for the job, is equal for both men and women (p = 0.5). Note that this algorithm does not adhere to predictive equality, as the probability of being hired, given not being qualified for the job is 0.5 for women, and 0 for men.

Candidate	Predicted outcome $\hat{Y}$	True outcome Y
Woman 1	hired	qualified
Woman 2	not hired	qualified
Woman 3	hired	not qualified
Woman 4	not hired	not qualified
Man 1	hired	qualified
Man 2	not hired	qualified
Man 3	not hired	not qualified
Man 4	not hired	not qualified

Lastly, the third notion of separation, *equalized odds*, combines the two previously described definitions, and is, therefore, most suitable in situations in which fairness is both sensitive to false positives and false negatives:

Equalized odds: a classifier satisfies equalized odds when the true positive rate and the false positive rate are equal for both the protected and the unprotected groups [52].

The recruitment prediction model would satisfy equalized odds when the probability of correctly hiring qualified candidates and the probability of incorrectly hiring unqualified candidates are both the same for female and male applicants:

$$P(\hat{Y} = 1|Y = 1, S = female) = P(\hat{Y} = 1|Y = 1, S = male)$$

$$P(\hat{Y} = 1|Y = 0, S = female) = P(\hat{Y} = 1|Y = 0, S = male)$$

Table 4 shows an example of a recruitment algorithm that satisfies equalized odds.

Table 4: Example of a recruitment algorithm adhering to equalized odds. This algorithm adheres to equalized odds as both (1) the probability of being hired, given being qualified for the job, is equal for both men and women (p = 1), and (2) the probability of being hired, given not being qualified for the job, is equal for both men and women (p = 0.5).

Candidate	Predicted outcome $\hat{Y}$	True outcome Y
Woman 1	hired	qualified
Woman 2	hired	qualified
Woman 3	hired	not qualified
Woman 4	not hired	not qualified
Man 1	hired	qualified
Man 2	hired	qualified
Man 3	hired	not qualified
Man 4	not hired	not qualified

Sufficiency Lastly, a fairness criterion can also be conditioned on the predicted outcome instead of the true outcome. In that case, the criterion belongs to the category of **sufficiency**. Sufficiency requires equal true outcomes over people that are given equal predictions, regardless of the sensitive attribute.

Hence, again, for sufficiency to hold, the true outcome should be available. The two most discussed notions in this category are *predictive parity* (a.k.a. *outcome test*) and *calibration* (a.k.a. *test fairness*) [14].

Predictive parity: a classifier satisfies predictive parity when the positive predictive value (PPV, a.k.a. *precision*), is equal for both the protected and unprotected groups [14].

In our running example, this means that the recruitment prediction model would satisfy predictive parity if, for both female and male job candidates, the probability that a candidate is actually qualified for the job, given that the model has hired the applicant, is equal:

$$P(Y = 1|\hat{Y} = 1, S = female) = P(Y = 1|\hat{Y} = 1, S = male)$$

Table 5 shows an example of a recruitment algorithm that satisfies predictive parity. Similar to predic-

Table 5: Example of a recruitment algorithm adhering to predictive parity. This algorithm adheres to predictive parity as the probability of being qualified for the job, given being hired, is equal for both men and women (p = 1). Note that this algorithm does not adhere to equality of opportunity, as the probability of not being hired, given being qualified for the job, is 0 for women and 0.5 for men.

Candidate	**Predicted outcome $\hat{Y}$**	**True outcome Y**
Woman 1	hired	qualified
Woman 2	hired	qualified
Woman 3	not hired	not qualified
Woman 4	not hired	not qualified
Man 1	hired	qualified
Man 2	hired	qualified
Man 3	not hired	qualified
Man 4	not hired	qualified

tive equality, the criterion of predictive parity does not take false negatives into account, and is therefore not suitable for a context in which equal false negatives are crucial for fairness [3].

A fairness criterion closely related to predictive parity is *calibration*. This criterion interprets the value of the predicted outcome $\hat{Y}$ as a probability score P of receiving the favorable prediction.

Calibration: a classifier satisfies calibration when for any probability score p between 0 and 1, both the protected and unprotected groups have the same probability of truly belonging to the favorable class [51].

Our example recruitment prediction model would hence be calibrated when, given any probability score p of being hired for the job, the probability of being qualified for the job would be the same, for both female and male job candidates:

$$P(Y = 1|P = p, S = female) = P(Y = 1|P = p, S = male) = \forall p \in [0, 1]$$

Table 6 shows an example of a recruitment algorithm that satisfies calibration. Calibration is a suitable fairness criterion in situations in which the threshold probability of receiving the favorable outcome varies. An example of such a situation is algorithmic loan approval: depending on the economic context, the acceptance score of granting a loan may increase or decrease [3].

Table 6: Example of a recruitment algorithm adhering to calibration. This algorithm adheres to calibration as for any theshold probability p of being hired, both men and women have an equal probability of being qualified for the job: for p = 0.9, this probability is 1, for p = 0.6, this probability is 0.5.

Candidate	True outcome Y	Threshold p
Woman 1	qualified	0.9
Woman 2	qualified	0.9
Woman 3	qualified	0.6
Woman 4	not qualified	0.6
Man 1	qualified	0.9
Man 2	qualified	0.9
Man 3	qualified	0.6
Man 4	not qualified	0.6

2.3 Impossibility Theory

Following the proliferation of research into mathematical criteria to define algorithmic fairness, several researchers have started to investigate their reciprocal mathematical relationships. This has exposed an important issue: there exist large trade-offs between several criteria, making them incompatible with each other. Therefore, it is impossible to satisfy all fairness criteria at once [48, 51, 53].

For example, Boracas et al. [48] describe that under mild assumptions, any two out of the three aforementioned categories of group fairness are mutually exclusive. For example, if the sensitive attribute and the true outcome are not independent, meaning that belonging to the unprotected group affects the statistics of the outcome, independence and sufficiency cannot be satisfied at the same time. So, to come back to our recruitment prediction model example, demographic parity and predictive parity could only hold at the same time, if both women and men would be exactly equally suitable for the job. Furthermore, Kleinberg et al. [53] prove that the criterion of equalized odds is inherently incompatible with the criterion of calibration, with the exception of two highly special cases: first, when the classifier achieves a perfect prediction, so one without false positives and false negatives, and second, when both the protected and unprotected groups have equal base rates, meaning they have the same percentage of true outcomes. These two conditions, however, are most often not the case. Lastly, Chouldechova [51] describes another incompatibility between two fairness criteria: she mathematically shows that predictive parity is in conflict with equalized odds unless the different groups have equal base rates.

Conflicts between fairness criteria such as those described above require practitioners, who are not yet always familiar with these topics, to make choices between the different criteria and their trade-offs. However, which choice to make is highly context-specific and can be a difficult task, given the subtle differences between the different criteria. Furthermore, which criterion to choose also depends on other factors, such as the availability of sensitive features, knowledge about the actual outcome label, and legal or organizational restrictions. More research is hence needed to understand the appropriateness of the different criteria in specific contexts and to understand their particular benefits and downfalls [12].

2.4 Technical approaches to mitigate algorithmic unfairness

To help researchers and practitioners understand and counter the fairness issues in their models, several toolkits have been developed. These toolkits largely focus on two main aspects: first, on measuring and evaluating biases in the model, and second, on mitigating these biases [46]. Table 7 shows a selection of the most prominently mentioned fairness toolkits.

Table 7: Overview of fairness toolkits

Toolkit	Main Focus	Features
Google's What-If toolkit [56]	Visualization and exploration	Interactive, open-source web application that offers visualizations for model probing and exploration. Allows users to compare two models, change feature values and evaluate performance and fairness metrics, with minimal coding. Includes a data point editor, which allows modifying the data and comparing counterfactual data points.
IBM Fairness 360 [41]	Bias mitigation	Python toolbox for detecting and mitigating biases in machine learning models. Offers a comprehensive set of fairness metrics including explanations. Provides a range of pre-processing, in-processing, and post-processing bias-mitigation algorithms.
Aequitas [57]	Fairness auditing	Open source bias and fairness bias audit report toolbox that automatically flags fairness issues. Includes a 'fairness tree' for helping users to choose the right fairness metric. Does not provide any algorithms to mitigate biases.
Microsoft Fairlearn [58]	Exploration and bias mitigation	Python library for assessing and improving fairness. Consists of two components: an interactive dashboard that enables users to try out multiple parity-based fairness metrics, and several bias mitigation algorithms to reduce unfairness.
FairML [59]	Fairness auditing	Open-source, end-to-end Python toolbox for auditing prediction models. Quantifies the effect of different input attributes, which can subsequently be used to investigate the fairness of the model.

However, despite these promising contributions for detecting biases and designing fairer models, two important concerns have arisen in the literature. First, there seems to be a disconnect between the practical application of these tools, and the fairness researchers constructing them. Factors such as context-dependence, application-dependence, and practitioners being unaware of the exact workings and possibilities of these tools, result in a lack of employment of them [12, 54]. Second, all of the aforementioned fairness criteria and fairness tools take a purely technical perspective on the topic of fairness and algorithmic bias. However, multiple authors state that more emphasis on the social, human side of fairness is needed: in order to develop fair AI, it is essential to know what humans perceive as fair and to acknowledge that fairness is not merely a technical construct [1, 12, 55].

Related work on human perceptions of algorithmic fairness

In this section, the empirical literature on human perceptions of algorithmic fairness will be discussed. First, Section 3.1 will analyze two predictors of perceived algorithmic fairness: factors influencing perceived algorithmic fairness, and human attitudes toward algorithmic fairness. Next, in Section 3.2, insights from a branch of psychology termed *organizational justice* will be drawn, to systematically describe the components of algorithmic fairness and link these to the existing empirical literature on fairness perceptions. Altogether, this section will aim to give a comprehensive overview of algorithmic fairness from a human perspective.

3.1 Human predictors of perceived algorithmic fairness

A considerable amount of fairness research investigates the influence of human (socio-)demographic factors, such as education level, age, or race, on the fairness judgment of algorithmic decision-making systems. Often, these are online, crowdsourced, survey-based studies in which participants have to rate the fairness of machine learning models in different contexts. A commonly used context is a criminal risk prediction context, in which participants have to judge the fairness of algorithms that estimate the risk of a defendant's criminal recidivism [24, 25]. By comparing these fairness ratings and taking into account the inter-individual differences between the participants, these studies look for correlations between certain (socio-)demographic factors and algorithmic fairness perceptions. In Table 8, some of the most important findings of these studies are stated. It is worth mentioning that these studies often present contradictory findings.

Furthermore, a number of empirical studies investigate different aspects of human attitudes toward the fairness of algorithmic decision-making. For example, some studies investigate whether participants' attitude toward the fairness of algorithmic decision-making differs per context [20, 21], or whether participants' attitudes toward the fairness of algorithmic decision-making differs from their attitudes toward the fairness of human decision-making [18]. In Table 9, some examples of these studies are shown. Again, these studies offer mixed results: for example, where Araujo et al. [21] conclude that people perceive automated decision-making in a criminal justice context as fair, Wang [20] finds opposite results.

In a critical paper, Dasch et al. [17] comment on some of the above-discussed papers by stating that to successfully understand and assess human perceptions of algorithmic fairness, an interdisciplinary approach, which incorporates insights from statistics and psychology, is needed. They stress that it is necessary to use the proper statistical methods and to be sensitive to psychological phenomena, such as framing, to derive reliable and generalizable results. Not only is more research needed to fully understand people's attitudes towards, and comprehension of, AI fairness, but it is also crucial to design these experiments in a reliable, replicable way, in order to be able to draw any solid conclusions. Moreover, Starke et al. [1] interpret the above-discussed inconsistencies in the literature on human fairness perceptions by arguing that "some of the inconclusiveness of the empirical results can also be attributed to the lack of coherent theoretical frameworks for perceived algorithmic fairness". Therefore, the following section will take a step toward adopting such a coherent framework by further outlining the concept of

perceived algorithmic fairness and dividing it into several components.

Table 8: Studies into human factors influencing perceived algorithmic fairness

Factor	Study	Algorithmic Context	Findings
Computer literacy	Wang et al. (2020) [19]	MTurk Master qualification algorithm	Positive correlation between computer literacy and perceived algorithmic fairness
	Pierson (2017) [24]	Recommendation algorithm, criminal risk prediction algorithm	Lecture about algorithms increases perceived algorithmic fairness
Education level	Wang et al. (2020) [19]	MTurk Master qualification algorithm	No significant effect
	Grgíc-Hlaca et al. (2020) [25]	Criminal risk prediction algorithm	No significant effect
	van Berkel et al. (2021) [60]	Criminal risk prediction algorithm, loan prediction algorithm	Negative correlation between education level and perceived algorithmic fairness
	Helberger et al. (2020) [18]	No specific context mentioned	Positive correlation between education level and perceived algorithmic fairness
Gender	Wang et al. (2020) [19]	MTurk Master qualification algorithm	No significant effect
	Grgíc-Hlaca et al. (2020) [25]	Criminal risk prediction algorithm	No significant effect
	Pierson (2017) [24]	Recommendation algorithm, criminal risk prediction algorithm	Women perceive gender inclusion in machine learning models as less fair than men
	van Berkel et al. (2021) [60]	Criminal risk prediction algorithm, loan prediction algorithm	Women associated with lower perceived algorithmic fairness levels
	Helberger et al. (2020) [18]	No specific context mentioned	No significant effect
Age	Wang et al. (2020) [19]	MTurk Master qualification algorithm	No significant effect
	Grgíc-Hlaca et al. (2020) [25]	Criminal risk prediction algorithm	No significant effect
	Helberger et al. (2020) [18]	No specific context mentioned	Negative correlation between age and perceived algorithmic fairness
Race	Wang et al. (2020) [19]	MTurk Master qualification algorithm	No significant effect
	Grgíc-Hlaca et al. (2020) [25]	Criminal risk prediction algorithm	No significant demographic effect, but men perceive using race as a feature as more fair than women
Political ideology	Grgíc-Hlaca et al. (2020) [25]	Criminal risk prediction algorithm	Liberals perceive gender- and race inclusion as less fair than conservatives

Table 9: Studies into human attitudes toward algorithmic fairness

Subject	Study	Findings
Algorithmic decision-making v.s. human decision-making	Helberger et al. (2020) [18]	54% of the respondents perceive AI decision-making as more fair compared to 33% of the respondents choosing a human decision-maker. The remaining 13% answers that they are both equally fair, or that this depends on the circumstance.
Fairness of automated decision-making in different contexts	Araujo et al. (2019) [21]	Respondents perceived AI decision-making as more fair in high-impact contexts (medicine and criminal justice) compared to a lower-impact context (media).
Fairness of automated decision-making in criminal justice system	Wang (2018) [20]	Respondents strongly disapprove of algorithms regarding fairness.
Comprehension of algorithmic fairness metrics	Saha et al. (2020) [61]	Participants' comprehension is lower for both equal opportunity and equalized odds, compared to demographic parity. A higher education level is a strong predictor for comprehending the metrics. Participants with a higher comprehension score tend to perceive the models as less fair.

3.2 Perceived algorithmic fairness: drawing insights from organizational justice theory

What it means for a decision to be fair, and what humans perceive as fair, have long been questions of interest in many fields, including philosophy, law, anthropology, neuroscience, and psychology [55]. When empirically investigating fairness perceptions, therefore, it is useful to draw insights from fields such as these. Much of the current literature on perceived algorithmic fairness takes inspiration from a branch of psychology termed *organizational justice* [1, 12]. Since this research area systematically describes the different components of perceived fairness, it can provide a solid foundation for how different aspects of mathematical algorithmic fairness can be studied, and connected to perceived algorithmic fairness.

In a seminal paper about fairness in organizations, Greenberg [30] introduces the concept of organizational justice: a research area concerned with fairness (or justice) perceptions of decisions made about employees in workplace settings, investigating the impact of perceived fairness on the functioning of businesses. In organizational justice literature, it is demonstrated that increased perceived fairness in organizations is related to beneficial outcomes, such as increased employee satisfaction and greater organizational commitment [31]. On the other hand, the literature on organizational justice also shows that when employees think they receive unfair treatment, they often react by decreasing their contributions to the company or by resigning from their position [62]. These findings are interesting for algorithmic fairness researchers: if insights from organizational justice can be connected to fairness

in algorithmic decision-making, these insights could be leveraged to gain a better understanding of, and potentially increase, perceived algorithmic fairness. Similar to algorithmic decision-making, organizational justice focuses on decisions made about others in a hierarchical setting, making this area a suitable source of inspiration for studying perceived algorithmic fairness [16]. Considering that organizational justice mainly focuses on the perceived fairness of employees in work environments, it may particularly be interesting to apply this theory to algorithmic hiring, the context I address in this thesis.

Within organizational justice theory, four different dimensions of perceived fairness have been validated: *distributive fairness*, a type of fairness concerned with the outcome of decisions, *procedural fairness*, a type of fairness concerned with the process of decision-making, *informational fairness*, a type of fairness concerned with the explanations and information provided for decisions, and *interpersonal fairness*, a type of fairness concerned with how decision-subjects are being treated. The last two of these dimensions, informational and interpersonal fairness, are often referred to together as *interactional fairness* [31].

While the different fairness dimensions in organizational justice theory are correlated and show some overlap, in the following sections, they will first be described separately. Furthermore, per dimension, related work will be discussed that focuses on one of these different dimensions of perceived fairness in an algorithmic context.

3.2.1 Distributive fairness

Distributive fairness (or outcome fairness) refers to the fairness of outcome distributions. It is based on norms for outcome allocation, such as equality (outcomes should be distributed equally amongst everyone) and equity (opportunities should be distributed equally based on everyone's circumstances) [1, 5, 31]. Distributive fairness can be assessed by looking at how outcomes are distributed, or how resources are allocated, across group members. This assessment is influenced by several factors, such as whether the decision is in favor of the individual assessing its fairness, or which resources are being distributed [19].

Related work on distributive algorithmic fairness

Robert et al. [62] note that in the literature on the perceived fairness of algorithmic decision-making, distributive fairness is the most commonly discussed category of perceived algorithmic fairness. One possible explanation for this finding could be the fact that many statistical fairness criteria focus on outcome distributions (for example, demographic parity requires the percentage of favorable outcomes to be equal across groups) . Dolata et al. [55] refer to this conclusion as the *distributiveness assumption*: the assumption that all fairness concerns can be represented as an outcome distribution problem. However, much of the work on perceived distributive fairness in machine learning focuses on basic fairness concepts, such as equality and equity [1]. Only a handful of studies focus on the distributive fairness of particular mathematical fairness criteria specifically [5, 22, 23, 63]. Here, I will discuss these studies in further detail.

First of all, Morse et al. [5] investigate the distributive fairness of several popular mathematical fairness criteria: fairness through unawareness, demographic parity, accuracy parity, equality of opportunity, and equalized odds. They do so, by categorizing these five criteria along the extent to which they accomplish the goals of distributive fairness. By reflecting on the different criteria and analyzing related literature, they reach the following classification:

- *Low level of distributive fairness*: fairness through unawareness
  - Morse et al. argue that fairness through unawareness cannot always ensure a fair outcome distribution, given the possibility of variables correlating with the ignored protected attributes.
- *Moderate level of distributive fairness*: demographic parity, accuracy parity

  - A moderate level of distributive fairness is given to demographic parity and accuracy parity since these notions require equal acceptance or accuracy rates across subgroups: they, therefore, do not take subgroup differences into account.

- *High level of distributive fairness*: equalized odds, equality of opportunity

  - Morse et al. assign the highest level of distributive fairness to equalized odds and equality of opportunity, given the fact that these definitions ensure more equitable outcomes by taking into account subgroup differences.

Although this classification provides a good starting point for investigating how different fairness criteria are related to distributive fairness aspects, it does not take into account how people actually perceive the distributive fairness of these criteria. The studies by Srivastava et al. [22] and Harrison et al. [23] address this question, by investigating how humans perceive the fairness of several group-level mathematical fairness criteria.

By running an experiment in which crowdworkers have to choose between a succession of pairs of machine learning model outcomes, Srivastava et al. [22] try to identify the mathematical fairness criterion that best captures human perceptions of fairness in a skin cancer risk prediction context and criminal risk prediction context. The outcomes of the models are visualized as 10 images of both Black and White males and females, accompanied by green and red visualizations of the algorithm's predicted outcomes and the true outcomes (whether a person has a low or high risk of skin cancer, or whether a criminal re-offends or not). In a series of 20 comparisons, generated by an adaptive algorithm, participants have to decide which of the two models is more discriminatory. In both contexts, Srivastava et al. find that participants prefer demographic parity over more complicated definitions, such as error parity and equal false positive rates. This finding suggests that humans exhibit a preference for fairness criteria that are more simplistic in nature.

However, Harrison et al. [23] draw different conclusions. They perform a between-subjects survey-based experiment in a bail decision-making context, in which they let participants judge two models with pairwise trade-offs between accuracy, outcomes, false positive rates, and the consideration of race. By doing so, they investigate which fairness criterion is preferred by the participants. Two interesting preferences are identified: first, subjects favor equalizing the false positive rate over equalizing the accuracy across groups. Second, subjects also favor equalizing the false positive rate over equalizing the percentage of favorable outcomes (i.e., having demographic parity) across groups. This latter result is contrasting with that of Srivastava et al. [22]. These findings hence raise several questions, such as what the effect of showing participants different kinds of visualizations is, and how the way questions are asked or information is provided to participants influences the outcomes.

Lastly, the study by Saxena et al. [63] focuses on the perceived distributive fairness of individual fairness criteria, rather than group-level fairness criteria. In a loan decision-making scenario in which participants have to judge the fairness of dividing a loan between individuals with disparate repayment rates, Saxena et al. assess crowdworkers' fairness perceptions of three different criteria: treating individuals that are similar in a similar way, never favoring worse over better individuals and selecting individuals proportional to their merit. They show a significant overall preference for the last criterion, which is similar to calibration. Although this experiment concerns individual fairness instead of group-level fairness criteria, and concerns another kind of decision-making scenario, this preference of dividing the outcome proportionally is still somewhat contrary to the findings of Srivastava et al. [22], who find that participants tend to prefer dividing the favorable outcome equally across groups. Both authors conclude that more research into people's assessment of fairness criteria is needed for a better comprehension of human perceptions of distributive algorithmic fairness.

To conclude, the above-described studies report contrasting results with regard to which mathematical fairness criterion best captures human perceptions of distributive fairness. Therefore, this research inconsistency will be further investigated in this thesis' user study.

3.2.2 Procedural fairness

The second dimension of perceived fairness, procedural fairness (or process fairness), refers to the fairness of the procedures that are needed to arrive at a decision or to reach an outcome. Instead of looking at the outcome itself, procedural fairness concerns the fairness of *how* a decision is taken. This is a particularly interesting dimension of perceived fairness, as it is of great importance in supporting a decision: it has been shown that people mostly rely on procedural fairness when it is uncertain whether to trust the decision-maker [64]. Six components of procedural fairness are distinguished by Leventhal [65]: consistency, bias suppression, correctability, ethicality, representativeness, and accuracy. Correctability, for example, reflects the control of individuals over the decision process: for example, if the decision subject has the possibility to correct faulty outcomes, this is likely to improve procedural fairness [5] [62].

Related work on procedural algorithmic fairness

So far, in the context of the perceived fairness of algorithmic decision-making, procedural fairness has received relatively little attention compared to distributive fairness [29]. This may be attributed to the fact that algorithms are frequently referred to as a *black box*, as it is not always entirely clear how an algorithm has exactly arrived at its decisions. As a result, algorithmic decision-making processes sometimes lack transparency. Nevertheless, it is possible to examine the fairness of specific aspects of an algorithmic decision-making process. For example, one could investigate the fairness of the features an algorithm uses as input. In the following, I will discuss three studies that address procedural algorithmic fairness: one about the different components of procedural fairness [5], one about the procedural fairness of using specific features in a model [27], and one about the effect of outcome control and transparency in the decision-making process [19].

Besides categorizing mathematical fairness criteria along their extent of distributive fairness, Morse et al. [5] also investigate the procedural fairness of these criteria. They compare fairness through unawareness, demographic parity, accuracy parity, equality of opportunity, and equalized odds along the six components of procedural fairness described by Leventhal [65]. They argue that while all of these criteria score high on consistency (since algorithms are able to surpass human consistency, they note), and all of them score low on correctability (they do little to correct faulty decisions), differences between the criteria exist between the other four components: accuracy, ethicality, representativeness, and bias suppression. For example, they contend that given the problem of proxy attributes, fairness through unawareness scores lower on each of these four components compared to the other criteria, possibly leading to negative fairness perceptions. Parity definitions score higher on representativeness compared to other criteria, since by ensuring equal success rates between groups, they can explicitly address concerns related to representativeness. Furthermore, they argue that equal opportunity and equalized odds score higher on bias suppression than the other criteria, given their targeted approach and strict rules to prevent preferential treatment. By relating these fairness criteria to the different components of procedural fairness, Morse et al. [5] provide directions for choosing the right criterion per situation and provide a fundament for better understanding and assessing the procedural fairness of these criteria: they, for example, reason that equality of opportunity and equalized odds are criteria with a high level of procedural fairness.

Grgic-Hlaca et al. [27] take a different approach to investigate procedural algorithmic fairness: they seek to identify feature properties that influence the perceived fairness of using certain attributes as input for an algorithmic decision-making model. They describe eight different latent properties of features that influence human fairness perceptions and investigate participants' assessments of these properties. By letting participants judge the fairness of these properties, they reach the following order of properties (from most fair to least fair): relevance, causes outcome, reliability, privacy, volitionality, causes vicious cycle, causes disparity in outcomes, and lastly, caused by sensitive group membership. As some of these feature properties, such as relevance or privacy, are unrelated to discrimination, Grgic-Hlaca et al. conclude that procedural unfairness concerns reach far beyond discrimination only and that therefore, other feature properties such as privacy, should also be taken into account when assessing algorithmic fairness.

Lastly, Wang et al. [19] investigate another aspect of procedural fairness: they hypothesize that perceived procedural fairness is higher when an algorithmic decision-making process is transparent (i.e. when public information about the workings of the model is provided), and includes human involvement. They survey participants about the fairness of a hypothetical machine learning model deciding whether Amazon Mechanical Turk workers will earn a Master's Qualification, with different kinds of development procedures (changing levels of transparency and human involvement). As they do not find any strong relationships between these procedures and perceived fairness, they reject their initial hypotheses. As a possible explanation for these findings, they contend that laypeople may not directly understand the relationship between fairness and development procedures: they, therefore, note that more information and explanations about algorithmic decision-making may increase perceived procedural fairness, pointing out to aspects of interactional fairness.

To conclude, in the user study that will be conducted in this research, three important insights from the above-described studies regarding procedural fairness can be taken: the finding that some mathematical fairness criteria such as equality of opportunity score high on different aspects of procedural fairness, the idea that feature properties play an important role in assessing perceived algorithmic fairness, and finally, the suggestion that providing explanations about decisions may help in assessing the procedural fairness of an algorithm.

3.2.3 Interactional fairness

The last aspect of perceived fairness, interactional fairness, can be subdivided into interpersonal fairness and informational fairness. Since these two are closely related, they are often taken together. Interpersonal fairness refers to the extent to which decision-makers communicate their choices with honesty, dignity, and respect to the people affected by their decisions. Informational fairness refers to providing sufficient information and giving truthful explanations about decision procedures. It is concerned with presenting people with adequate information about the process of how a decision is reached and is therefore closely related to procedural fairness [16, 31]. In an organizational justice setting, an example of interactional fairness is providing employees explanations for layoff decisions: it has been shown that if employees receive honest, thorough, and accurate explanations when being fired, they perceive these decisions as significantly more fair [28].

Related work on interactional algorithmic fairness

Multiple researchers investigate the effect of explanations for decisions, an important aspect of interactional fairness, on the perceived fairness of algorithmic decision-making. Here, three of these studies will be discussed.

First of all, Binns et al. [16] perform an online user study using fictional fairness scenarios and explanations in an insurance context. They investigate the influence of four different explanation styles: input-influence-based explanations (reporting the importance of different input features), demographic-based explanations (reporting aggregate statistics of the outcomes of people in the same demographic classes), case-based explanations (reporting similar cases, along with their outcomes) and sensitivity-based explanations (reporting how feature values have to change to get a different outcome). When participants are presented with multiple explanations simultaneously, case-based explanations result in a significantly lower fairness perception compared to other explanation styles.

In a study on the effect of different explanation styles on the perceived fairness of decisions made by a criminal risk prediction algorithm, Dodge et al. [26] build on the research of Binns et al. [16]. They use the same four explanation styles, but instead of manually creating them, they automatically generate the explanations. Besides, instead of creating a fictional scenario, they use the COMPAS data set and implement a logistic regression model. After identifying cases with unfair treatment and sampling different cases for a user study, they conduct an online user study in which they investigate the perceived

fairness of the four different explanation styles. Similar to the research of Binns et al. [16], they find that case-based explanations are generally seen as the least fair explanation type. Furthermore, they find that input-influence and demographic-based explanations increase the participant's comprehension of the model, resulting in higher fairness perceptions.

A recent study by Angerschmid et al. [66] furthermore examines the effect of explanations on human trust and perceived algorithmic fairness. They perform an online user study in a health-insurance and medical decision-making context, in which different AI-informed decision-making scenarios are simulated and accompanied by either an example-based explanation (similar to a case-based explanation), a feature importance-based explanation (similar to an input-influence explanation), or no explanation (the control group). The results indicate that explanations in general give a significantly increased level of both trust and perceived algorithmic fairness. Furthermore, the authors show that in unfair scenarios, feature importance-based explanations lead to higher perceived fairness levels, compared to example-based explanations. Regarding human trust, they do not find differences between the two explanation styles.

Although the three studies described above investigate the effect of explanations on perceived fairness in different contexts, two important insights can be drawn: first, both Binns et al. [16] and Dodge et al. [26] find that case-based explanations result in a lower level of perceived fairness, compared to other explanation styles. This is an interesting finding, as it could be argued that case-based explanations align somewhat with the idea of using of individual, similarity-based fairness criteria. Second, both Angerschmid et al. [66] and Dodge et al. [26] find a positive effect of explanations that describe the influence of the different input features used in the model. This latter insight will be used in the user study that will be conducted in this thesis, by providing the participants with feature importance explanations about machine learning models.

To summarize the insights from this section, Table 10 presents an overview of the different dimensions of fairness discussed in organizational justice theory. Per dimension, it provides an example question addressing the specific type of fairness in both a human– and algorithmic decision-making context.

Table 10: Overview of different fairness dimensions in organizational justice theory

Name	Definition	Human example	Algorithmic example
Distributive fairness	Fairness regarding the distribution of outcomes	Is the amount of money paid by a boss to male and female workers equal?	Does an automated hiring algorithm accept as many female and male candidates?
Procedural fairness	Fairness regarding the process utilized to achieve an outcome	Are the punishments for arriving late to work consistently in all cases?	Does an algorithm that judges cover letters take into account race?
Interactional fairness	Fairness regarding the information provided about decision procedures	Does a manager provide truthful and respectful explanations about a layoff decision?	Can outcomes of an automated hiring algorithm be explained in an understandable manner?

Methods

This section will provide an overview of the data and various research methods used in this thesis. The methodology of this thesis consisted of two stages. In the first stage, machine learning models were created and tuned to adhere to different fairness criteria, as explained in Section 4.2. The second stage consisted of a user study that made use of the results of these models. Section 4.3 will describe the design and procedures of this user study.

4.1 Data

For this study, I used a new, publicly available data set in the recruitment domain, that is published in 2022 on Kaggle [1]. The data set is created by Sieuwert van Otterloo, a Dutch AI researcher at the Vrije Universiteit Amsterdam and Utrecht University of Applied Sciences. The data set contains information on the recruitment decisions of four different hypothetical consulting companies. Of each company, data from 1000 candidates are provided, amounting to 4000 instances in total. The data set does not contain any missing values. An overview of the thirteen attribute characteristics in this data set can be found in Table 11. The data set is specially designed to mimic realistic recruiting data and to gain a deeper understanding of AI fairness. This is done by including several variables that can be analyzed for bias: age, gender, sport, and nationality. As the data description informs, including any of these attributes as an input variable can lead to a biased model. The data set is generated in a structured manner, by including hidden variables such as personality type. The data is fictional and not based on the hiring decisions of real companies. No results regarding classifier performance on this data set are published. However, the data description suggests that if all indicators are used, prediction models such as logistic regression, decision trees or neural networks should give good results.

Because hiring is a domain in which algorithmic decision-making is often used and does likely speak to the imagination of laypeople, I choose to focus on this domain specifically. While this area has seen increased interest in the integration of AI-enabled software, it has also witnessed raising concerns about the potential of AI to perpetuate or exacerbate existing biases [32, 34]. The use of algorithms in the hiring domain has already led to fairness issues in the past, with Amazon's biased recruitment model as a notorious example [8]. As a result, it is classified as a high-risk area in the proposed EU AI act [35]. Moreover, there is no universal agreement on how fairness should be formalized in algorithmic hiring: for instance, certain recruitment algorithms proactively aim to increase diversity when ranking job candidates, while others do not [36]. As research has demonstrated that fairness perceptions during a hiring process play a critical role in job satisfaction, performance, and the relationship between employers and employees, obtaining insights into the perceived fairness of algorithmic hiring is of particular importance [37]. By doing research into algorithmic fairness perceptions in this specific domain, I, therefore, hope to contribute to the field of AI Fairness.

[1] https://www.kaggle.com/datasets/ictinstitute/utrecht-fairness-recruitment-dataset

Table 12: Top six most important attributes of recruitment prediction model

Attribute	Importance
Languages	+ + + + +
Highest degree	+ + + + +
University grade percentage	+ + ++
Debate club	+ + ++
Exact study	++
Gender	++

nationality and age. In Appendix A.1, these count plots are showed. As the purpose of the model implementation stage was to create models to illustrate fairness issues with, I chose to proceed with the data subset with the largest visible bias. In this case, this considered a data subset with a large difference in selection rates between male and female candidates: the plots from the second company showed that the hiring rate of male candidates was larger than the hiring rate of female candidates. This bias was confirmed by calculating the disparate impact ratio: dividing the proportion of female candidates receiving the favorable outcome by the proportion of male candidates receiving the favorable outcome, led to a ratio of 0.3. In Appendix A.2, the count plots of the distribution of the target variable amongst the different categorical features in this data set are showed. As a next pre-processing step, the resulting data set was split up into a train set of size 750 and a test set of size 250. Lastly, the numerical features were scaled using Sklearn's MinMaxScaler[5] and the categorical features were encoded using Sklearn's OneHotEncoder[6].

4.2.2 Recruitment prediction model

The model implemented was a binary classifier, predicting whether a candidate in the recruitment data set is hired by the company or not. I chose to use a simple logistic regression model, for two reasons: first, to allow for a straightforward investigation of the model's feature importance, and second, since many current algorithmic decision-making systems rely on regression models [26]. The model was implemented using SKlearn's Logistic Regression [7]. Using SKlearn's default parameters, the accuracy of this model on the training data and test data was 88.4 % and 87.6 % respectively. As these accuracies were quite high already, and the model was solely created to demonstrate fairness issues, I decided not to further tune any parameters.

To investigate the importance of the different features used by the model, the model coefficients were sorted in ascending order. Subsequently, all features were split up into 11 buckets. This was done to rank the features in an importance order ranging from − − − − − to + + + + +, where the more +'s or -'s means a candidate with that attribute is respectively more or less likely to be hired to the company. The features with the most +'s, that were used in the experiments following the model development stage, are reported in Table 12. Note that here, '*importance*' is defined as having a positive influence on receiving the favorable outcome, i.e, being hired by the company. Appendix A.3 shows an overview of all features and their belonging coefficients and importance buckets.

To investigate the fairness of the model, I used Microsoft Fairlearn [58]. This Python library includes several functions to compute fairness metrics of a model[8]. Because of the disparity in selection rates between male and female candidates in the data set, *Gender* was specified as the sensitive attribute to

[5] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
[6] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
[7] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[8] https://fairlearn.org/v0.5.0/api_reference/fairlearn.metrics.html

Table 11: Utrecht Fairness Recruitment data attribute characteristics

Attribute	Values	Type	Description
Gender	Female, male, other	Categorical	The gender of the candidate
Age	[21-32]	Numerical	The age of the candidate in years
Nationality	German, Dutch, Belgian	Categorical	The candidates' current nationality
Sport	Swimming, golf, running, cricket, chess, tennis, football, rugby	Categorical	The first/main sport the candidate listed on their CV
University grade percentage	[47-77]	Numerical	University grade (percentage) of the candidate
Debate club	True, False	Binary	Whether the candidate participated in a debating / social club
Programming experience	True, False	Binary	Whether the candidate has programming experience
International experience	True, False	Binary	Whether the candidate has international experience
Entrepreneurship	True, False	Binary	Whether the candidate has run their own company
Languages	[0-4]	Numerical	Number of additional languages spoken fluently by candidate
Exact study	True, False	Binary	Whether the candidate studied physics, engineering or any other science-oriented study
Highest degree	Phd, Bachelor, Master	Categorical	Highest completed degree
Decision	True, False	Binary	Whether the candidate was hired (the target class to be predicted)

4.2 Model implementation

Prior to conducting the user study, I created machine learning models to demonstrate and test various fairness aspects. This section describes the data pre-processing steps needed to create these models and explains the model implementation process. Implementation was done in a Jupyter Notebook environment[2] using Python 3.8.5[3]. The source code detailing the exact implementations can be found on Github [4].

4.2.1 Data preprocessing

As the recruitment data set contains data about four different consulting companies, as a first preprocessing step, the data set was split up into four separate subsets of size 1000. Each subset contained data from one individual company. I decided to drop the sports attribute from the data, because it seemed less relevant for analyzing fairness issues compared to the other features. Next, for all four subsets, count plots were created to show the distribution of the target attribute (i.e., being hired to the company or not) amongst the features that could be considered as sensitive attributes: gender,

[2] https://jupyter.org/
[3] https://www.python.org/downloads/release/python-385/
[4] https://github.com/GuusjeJuijn/fairness-perceptions

compute the fairness metrics with. A large difference in selection rates and a considerable difference in false negative rates between the different genders was found, as reported in Table 13. As out of the 250 candidates in the test set only 4 gender identities were classified as *other*, for simplicity, only the values for the *female* and *male* candidates are reported in this table. The complete results, however, can be found on GitHub.

Table 13: Fairness metrics by group of original recruitment prediction model, demographic-parity mitigated recruitment prediction model and equality of opportunity-mitigated recruitment prediction model

Model	Gender	Accuracy	Selection rate	False negative rate
Original	Female	0.92	0.12	0.33
	Male	0.85	0.47	0.16
	Difference	*0.07*	*0.35*	*0.17*
Demographic parity mitigated	Female	0.89	0.20	0.13
	Male	0.74	0.25	0.51
	Difference	*0.15*	*0.05*	*0.38*
Equality of opportunity mitigated	Female	0.93	0.16	0.13
	Male	0.85	0.47	0.16
	Difference	*0.08*	*0.31*	*0.03*

4.2.3 Bias mitigation

As a next implementation step, I applied bias mitigation to the logistic regression model. This was done using the ThresholdOptimizer algorithm from Microsoft FairLearn [9]. This postprocessing algorithm, introduced by Hardt et al. [67] , adjusts a learned classifier by applying group-specific thresholds, to satisfy a specified fairness constraint, with respect to a specified sensitive feature. The ThresholdOptimizer was applied twice to the raw logistic regression model: the first time to mitigate for demographic parity and the second time to mitigate for equality of opportunity.

Postprocessing for demographic parity and equality of opportunity specifically was done for several reasons. First of all, multiple studies suggest that both of these criteria are appropriate for algorithmic hiring, the context I focus on in my empirical study [5, 33, 36, 38]. Mitigating for demographic parity, moreover, allowed for further investigation of the results of Srivastava et al. [22], who found that lay people tend to have a preference for this criterion in different contexts. Besides, as demographic parity is often used in practice and relatively easy to understand, I considered this a suitable criterion for this study [61]. Since, according to Morse et al. [5], equality of opportunity scores high on procedural fairness, I considered this a second suitable criterion.

The accuracies and fairness metrics of the mitigated models are reported in Table 13. Mitigation for demographic parity led to a considerably lower difference in selection rates between the gender groups, compared to the original model. However, it led to a larger difference in false negative rates between the groups. Mitigation for equality of opportunity led to a considerably lower difference in false negative rates between groups, compared to the original model. However, this mitigated model still showed a large difference in selection rates between the groups.

[9] https://fairlearn.org/v0.8/user_guide/mitigation.html

The results of the raw logistic regression model, the classifier mitigated for demographic parity, and the classifier mitigated for equality of opportunity were subsequently used in the empirical study, as described in the following section.

4.3 Empirical Study

The second stage of this thesis' methodology consisted of a between-subjects online experiment in which participants judged the fairness of multiple hypothetical recruitment algorithms. These recruitment algorithms were based on the outcomes of the original and mitigated versions of the logistic regression classifier described in Section 4.2.2 and 4.2.3. Participants were asked to rate the fairness of these algorithms. The amount of information they received about these algorithms differed per group, as further explained in Section 4.3.1. The study was conducted in two phases. First, a pilot study amongst a group of twenty colleagues and acquaintances was performed, which led to valuable insights and improvements in the study design. Hereafter, a larger study was performed using crowdsourcing platform Prolific Academic[10] to address the main research questions. This study was distributed at the end of January 2023. Both studies were classified as low-risk by the Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing[11], requiring no further ethics or privacy assessment.

4.3.1 Study Design

The online survey was conducted using Qualtrics survey software[12]. Participants' fairness perceptions of several hypothetical recruitment algorithms were assessed using a direct measure based on Harrison et al. [23], asking *"Do you think this algorithm is fair?"*. To ensure that every participant had a similar definition in mind, they were provided with a fairness definition by Mehrabi et al. [39]: *"Fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits"*. Participants were asked to provide a judgment on a 7-point Likert scale, ranging from 1 ("not at all fair") to 7 ("completely fair"). Additionally, at the end of the survey, participants were asked to elaborate on the motivations behind their ratings through an open-ended query, asking *"In the previous questions, which factors did you consider most important in determining whether an algorithm was fair or unfair?"*. This question was asked to qualitatively investigate the rationales behind the respondents' fairness perceptions.

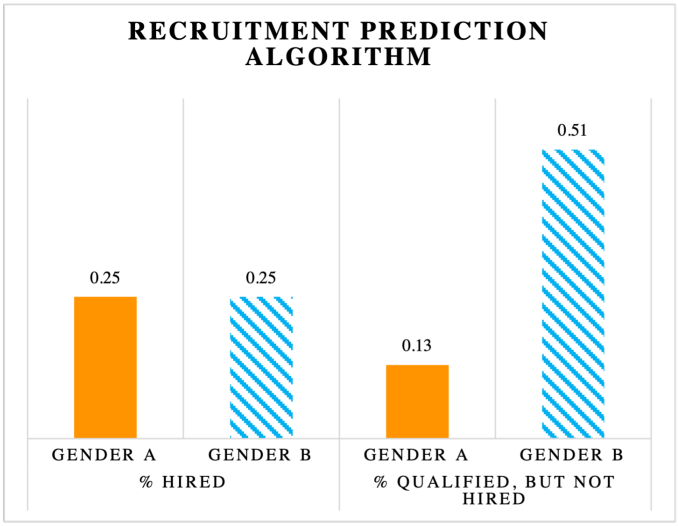
Each participant was presented with a selection of five out of nine different algorithms, of which the selection rates and false negative rates were based on the logistic regression models described in Table 13. Table 14 reports the selection rates and false negative rates of these nine algorithms. Every participant received an algorithm representing the original, unmitigated model. Moreover, every participant received two algorithms representing demographic parity – one perfectly following the criterion of demographic parity and one representing the mitigated model– and two algorithms representing equality of opportunity – again, one perfectly following the criterion of equality of opportunity and one representing the mitigated model. Participants were randomly assigned to either variant A or B of the algorithms representing demographic parity and equality of opportunity. I included two variants of these algorithms for two reasons: first, to broaden the investigation of participants' fairness perceptions, and second, to better investigate RQ3, which relates the differences in selection rates and false negative rates between groups to algorithmic fairness perceptions. Hence, including multiple variants broadened this correlational analysis.

Participants were divided into three groups. The amount of information participants received about these algorithms differed per group, based on the fairness components described in organizational justice

[10] https://www.prolific.co
[11] https://www.uu.nl/en/research/institute-of-information-and-computing-sciences/ethics-and-privacy
[12] https://www.qualtrics.com/nl/?rid=langMatch&prevsite=en&newsite=nl&geo=&geomatch=

theory. I chose to consider procedural and interactional fairness together, due to the strong connection and overlap between these two components. Next, I will describe these different groups.

Figure 1: Example outcome graph, representing distributive fairness, showed to each participant. On the left, the selection rates are shown. On the right, the false negative rates are shown. This algorithm adheres to demographic parity but not to equality of opportunity.



Group 1: distributive fairness The first group only received information about the distributive fairness of the algorithms. This information was visualized as a graph representing the algorithm outcomes, showing a pairwise trade-off between the selection rates and false negative rates between two gender groups. Instead of only showing one aspect of algorithmic fairness, by, for example, only showing the difference in false negative rates between groups, I chose to represent a more realistic real-world scenario by showing the trade-offs between different fairness criteria. By doing so, I drew inspiration from the work of Harrison et al. [23]. Furthermore, I explicitly chose to rename the two gender groups into *Gender A* and *Gender B*, to limit the effect of implicit biases regarding gender roles. An example of a graph representing distributive fairness is shown in Figure 1.

Group 2: distributive and procedural fairness, with sensitive attribute The second group not only received information about the distributive fairness of the algorithms, but also about the procedural fairness of the algorithms. Similar to Grgic-Hlaca et al. [27], I considered the features used by the algorithm as an important aspect of procedural fairness. Therefore, I visualized procedural fairness as a feature importance explanation. Similar to Dodge et al. [26], I presented the feature coefficients of the logistic regression models as strings of '+'s representing the relative importance of each feature. To limit the amount of information, only the top five most influential features were shown. For each of the algorithms, the feature importance graph stayed the same, as postprocessing does not change the model coefficients. Figure 2 displays the feature importance graph shown to the participants of group 2.

Group 3: distributive and procedural fairness, without sensitive attribute The information provided to group 3 was almost identical to that of group 2, except for a small change in the feature importance graph. In this group, the attribute *'gender'* was changed into a less sensitive attribute, with a similarly high feature coefficient: *'exact study'*. I included this group in the study to make sure that potential differences in fairness perceptions between the groups could not only be attributed to the use of the sensitive feature *gender* as an attribute.

Table 14: Different recruitment algorithms presented to participants. Each participant received 5 algorithms in random order. The algorithms of variant B were made by making slight adjustments to the selection rates of the algorithms of variant B. Each participant received the original graph, as well as the graphs of either variant A or B.

Graph	Selection rate	False negative rate
Original graph	Gender A: 0.12 Gender B: 0.47	Gender A: 0.33 Gender B: 0.16
Demographic parity-mitigated version A	Gender A: 0.20 Gender B: 0.25	Gender A: 0.13 Gender B: 0.51
Demographic parity version A	Gender A: 0.20 Gender B: 0.20	Gender A: 0.13 Gender B: 0.51
Equality of opportunity-mitigated version A	Gender A: 0.16 Gender B: 0.47	Gender A: 0.13 Gender B: 0.16
Equality of opportunity version A	Gender A: 0.16 Gender B: 0.47	Gender A: 0.13 Gender B: 0.13
Demographic parity-mitigated version B	Gender A: 0.30 Gender B: 0.38	Gender A: 0.13 Gender B: 0.51
Demographic parity version B	Gender A: 0.30 Gender B: 0.30	Gender A: 0.13 Gender B: 0.51
Equality of opportunity-mitigated version B	Gender A: 0.26 Gender B: 0.47	Gender A: 0.13 Gender B: 0.16
Equality of opportunity version B	Gender A: 0.26 Gender B: 0.47	Gender A: 0.13 Gender B: 0.13

4.3.2 Procedure

After signing a consent form, participants were shown an introductory text. This text can be found in Appendix A.4. The purpose of this text was to introduce the topic of algorithmic fairness, clarify the task, present the context, and demonstrate a sample graph to ensure that the participants could properly interpret the visual representations. Each participant was then randomly assigned to one of the three groups. Randomization was done automatically by Qualtrics. The participants were divided evenly across the groups to ensure that each group had an equal number of participants. Within each group, every participant was asked to rate the fairness of five different recruitment algorithms: one representing the original, unmitigated model, two adhering to demographic parity, and two adhering to equality of opportunity. These algorithms were presented in a randomized order to limit order effects. After these five questions, participants were asked to write down which factors they considered most important in their fairness analysis. Next, they were presented with the following three demographic questions: *"Do you have any experience with computer science and/or artificial intelligence?"*, *"To which gender do you most identify?"* and *"What is the highest level of education you have completed?"*. The survey ended with a message thanking the participants for their time and giving them a completion code to register their submission in Prolific. A graphical overview of the survey flow can be seen in Figure 3.

Figure 2: Feature importance graph showed to group 2, representing procedural fairness. The graph showed to group 3 was the same, except for the sensitive attribute 'gender' being changed for the non-sensitive attribute 'exact study'.

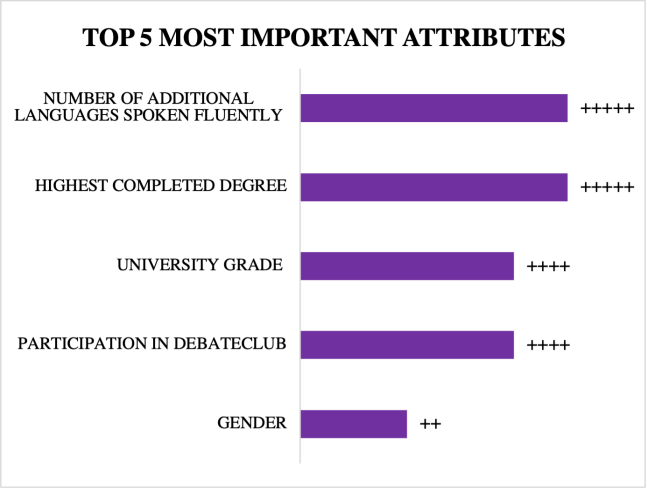
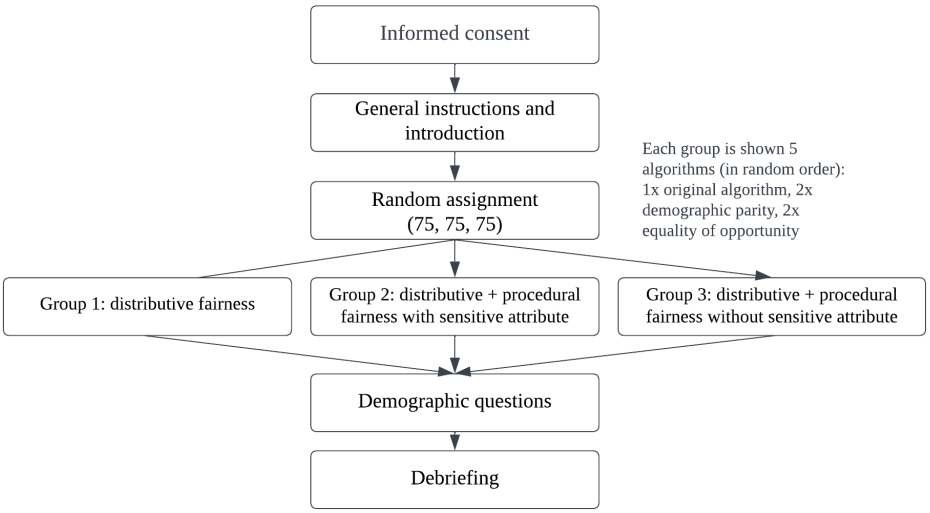


Figure 3: Experimental Flow



4.3.3 Participants

Participants were recruited through Prolific to take an online survey. To make sure all participants would understand the study properly, participants were pre-screened on age, education level, and first language: participants were required to be at least 18 years old, to have obtained at least a high school diploma, and to have English as a first language. Furthermore, only residents from the UK were selected, to minimize cultural biases among participants. Participants were compensated with an amount of £10,84 per hour, conforming to the minimum wage in the UK. On average, the survey took 4.2 minutes to complete. Data from 225 participants were collected. Table 15 summarizes their demographics. Participants' gender, age, and race/ethnicity were automatically collected by Prolific. The online survey specifically asked for their highest level of education obtained and whether participants had experience in computer science and/or AI.

Table 15: Participants' demographics

		Percentage (n=225)
Gender	Female	50%
	Male	50%
	Other	<1%
Age	18-30	33%
	30-45	35%
	45-60	22%
	60+	10%
Race/ethnicity	White	92%
	Asian	4%
	Mixed	3%
	Black	1%
Highest level of education obtained	High school diploma	54%
	Technical/community college	40%
	Undergraduate degree (BA/BSc/other)	5%
	Graduate degree (MA/MSc/Mphil/other)	<1%
	Doctorate degree (PhD/other)	<1%
Experience in computer science/AI	Yes	9%
	A little	13%
	No	78%

Results

This section will describe the results of the user study described in Section 4.3. First, a quantitative analysis of the results to research questions 1, 2, and 3 will be presented in Section 5.1 Next, Section 5.2 describes a qualitative analysis of the results of the open-ended question asked at the end of the survey. Finally, Section 5.3 presents a short demographical analysis of the impact of demographic characteristics on participants' fairness perceptions.

Results were analysed using Python 3.5.8[1], R 4.2.3 [2] and Microsoft Excel [68]. The fully anonymized survey data, as well as the source code detailing the exact analyses can be found on Github [3]. Appendix A.5 presents the survey data of three randomly selected participants, to clarify the survey format and to offer additional insights regarding participants' responses to the questions.

5.1 Quantitative Analysis

I will start with a quantitative analysis of the results of the empirical study. As all 225 participants, divided over three groups of 75 participants, rated five different algorithms, a total of 1125 perceived fairness scores were given. The average standard deviation of the scores given per participant was 1.1. The lowest standard deviation of the scores given per participant was 0, indicating a situation in which all five algorithms were rated with the same score. This was done by 17 participants. The highest standard deviation of the scores given per participant was 2.8, indicating a situation in which a participant gave a wide range of scores to the different algorithms.

In order to compare the average fairness perceptions of the different algorithms, I first tested whether there were any differences in scores for the algorithms that fully adhered to a criterion and for those that were mitigated for a criterion, as outlined in Table 14. Table 16 presents the result of this comparison. Interestingly, a larger number of participants gave a higher perceived fairness score to the algorithms mitigated for demographic parity, compared to the algorithms fully adhering to demographic parity. However, a Wilcoxon Signed Rank test (a non-parametric variant of the paired t-test) between the average scores for both algorithms, revealed no significant differences. Therefore, I decided to average their scores per person, to simplify the subsequent analysis. Most participants gave an equal perceived fairness score to the the algorithms mitigated for equality of opportunity, and the algorithms fully adhering to equality of opportunity. Again, a Wilcoxon Signed Rank test revealed no significant differences in average scores between both algorithms. Therefore, their scores were also averaged for the subsequent analysis.

5.1.1 RQ1

First, I investigated the effect of the type of information given about the algorithms on participants' fairness perceptions, to answer the first research question:

[1] https://www.python.org/downloads/release/python-385/
[2] https://cran.r-project.org/bin/windows/base/
[3] https://github.com/GuusjeJuijn/fairness-perceptions

Table 16: Comparison between average scores for algorithms mitigated for a fairness criterion and average scores for algorithms fully adhering to a fairness criterion. Rows add up to 225: the total number of participants.

Algorithm	Number of times the mitigated algorithm is rated equally fair as the algorithm fully adhering to the criterion	Number of times the mitigated algorithm is rated as less fair than the algorithm fully adhering to the criterion	Number of times the mitigated algorithm is rated as more fair than the algorithm fully adhering to the criterion
Demographic Parity	81	57	87
Equality of Opportunity	107	62	56

RQ1: *How do human fairness perceptions of a recruitment algorithm differ when only given information about the distributive fairness of the algorithm, compared to when given information about both the procedural fairness and the distributive fairness of the algorithm?*

For each of the three groups, I computed the average fairness perceptions of the original algorithm, the algorithms adhering to demographic parity, and the algorithms adhering to equality of opportunity.

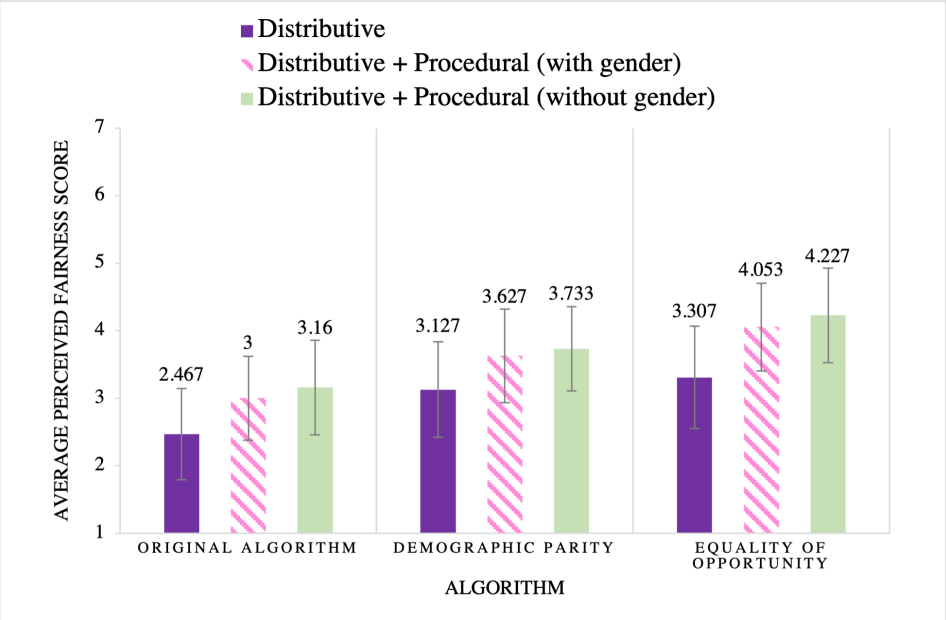
Figure 4 shows that participants who received information about both the distributive and procedural fairness of the algorithms (groups 2 and 3) consistently perceived the algorithms as fairer compared to participants who only received information about the distributive fairness of the algorithms (group 1). I observed this effect in both groups 2 and 3, although fairness perceptions were generally higher in group 3, in which the sensitive attribute gender was not included as a main attribute in the feature importance graph.

Table 17 reports the results of three Kruskal-Wallis H tests (a non-parametric variant of the ANOVA test to compare multiple groups), followed by multiple comparisons post-hoc Dunn tests, to test for significant differences between the three groups. The tests were performed separately for the different algorithms (the original algorithm, the algorithms adhering to demographic parity, and the algorithms adhering to equality of opportunity). Results indicated significant differences between groups 1 and 2, and groups 1 and 3, for all algorithms. Differences between groups 2 and 3 were not significant.

Table 17: Results of Kruskal-Wallis H tests and post-hoc Dunn tests to test for significant differences between the three groups. P-values are in italics if results are significant at α=0.05. Results of the Kruskal-Wallis H tests indicate that the average scores, for all algorithms, differ significantly across groups. Pairwise comparisons by Dunn's tests show that differences between groups 1 and 2, and 1 and 3, are significant at α=0.05. Differences between groups 2 and 3 are not significant.

Algorithm	Kruskal-Wallis H test		Dunn's Multiple Comparisons test		
			Groups 1-2	Groups 1-3	Groups 2-3
	H	p	p	p	p
Original	10.691	*0.005*	*0.009*	*0.003*	0.715
Demographic Parity	8.452	*0.014*	*0.044*	*0.005*	0.419
Equality of Opportunity	18.127	*<0.001*	*0.001*	*<0.001*	0.468

Figure 4: Average perceived fairness scores, on a 7-point Likert scale, of each of the three groups. Error bars indicate standard deviations. Bar graphs show that the group that only received information about the distributive fairness of the algorithms rated each of the three algorithms lower than the groups that also received information about the procedural fairness of the algorithms. In the group in which gender was not a main attribute, fairness perceptions were highest.



5.1.2 RQ2

Next, I investigated whether participants preferred either the algorithms adhering to demographic parity or the algorithms adhering to equality of opportunity, to answer the second research question:

RQ2: *How do human fairness perceptions of a recruitment algorithm differ depending on whether it adheres to demographic parity or equality of opportunity?*

For each of the three groups, I computed the average fairness perceptions of the algorithms adhering to demographic parity and the average fairness perceptions of the algorithms adhering to equality of opportunity. Figure 5 shows that across all three groups, participants tended to have a preference for the algorithms adhering to equality of opportunity. Table 18 reports the results of a Wilcoxon-Signed Rank test for each of the three groups. Results indicated that in groups 2 and 3, the average fairness perceptions of the algorithms adhering to equality of opportunity were significantly higher than the average fairness perceptions of the algorithms adhering to demographic parity. However, in group 1, these differences were not statistically significant.

Figure 5: Average perceived fairness scores, on a 7-point Likert scale, of the algorithms adhering to demographic parity and equality of opportunity. Error bars indicate standard deviations. Bar graphs show that across all three groups, algorithms adhering to equality of opportunity were rated higher compared to algorithms adhering to demographic parity.

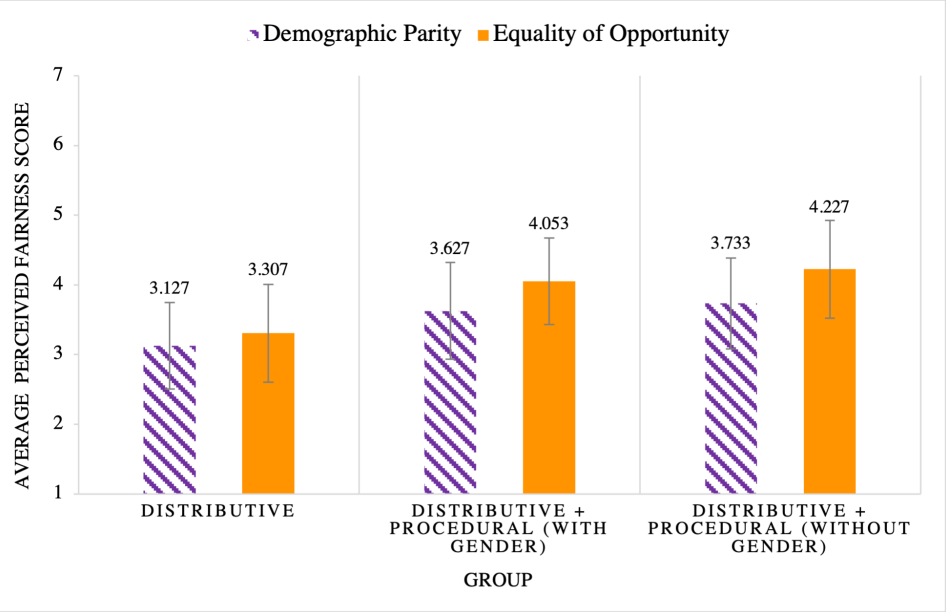


Table 18: Results of Wilxocon-Signed Rank tests to compare the mean perceived fairness scores of the algorithms adhering to demographic parity and the algorithms adhering to equality of opportunity, for each group. P-values are in italics if results are significant at α=0.05.

Group	W	P-value
Distributive	777.0	0.541
Distributive + Procedural (with gender)	601.5	*0.013*
Distributive + Procedural (without gender)	636.5	*0.016*

5.1.3 RQ3

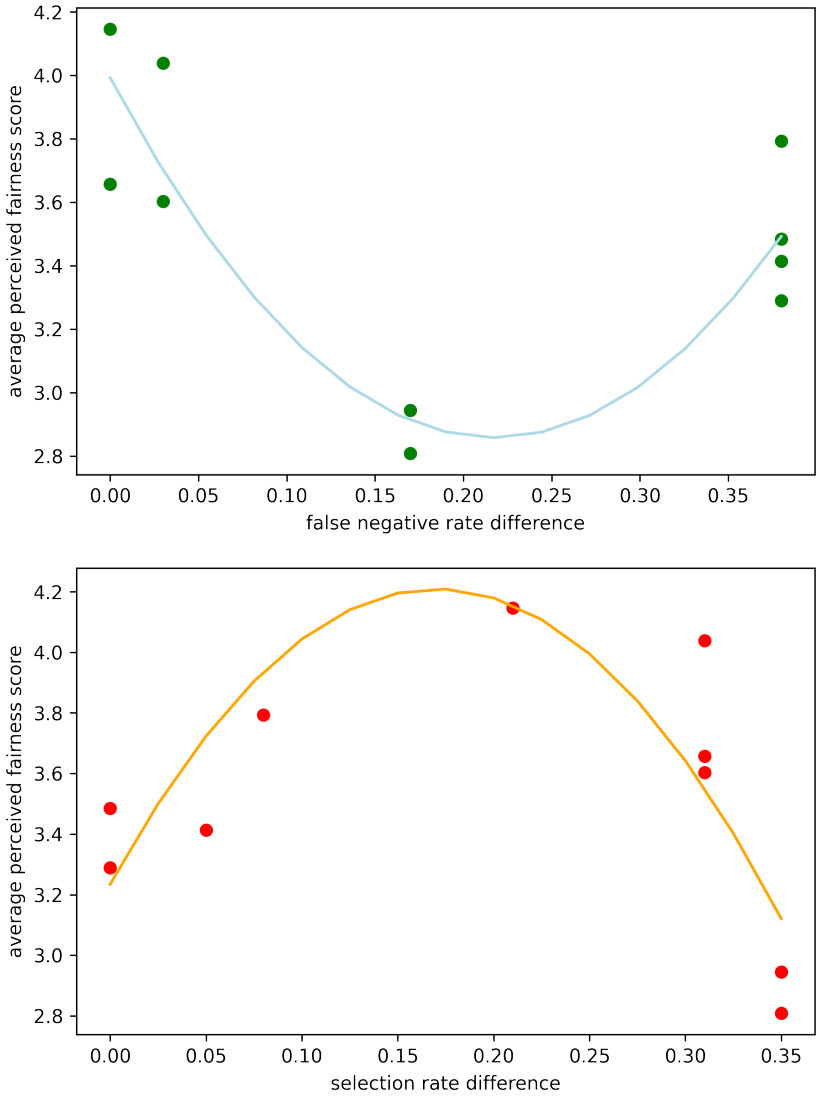
Thirdly, I investigated the relationship between fairness perceptions and two mathematical fairness criteria, to answer the third research question:

RQ3: *To what extent are the disparate impact score and false negative rate differences between groups of a recruitment prediction algorithm related to human fairness perceptions of it?*

I computed the Spearman correlation coefficient between the averaged perceived fairness scores of each of the nine different algorithms and, respectively, their belonging selection rate differences, and false negative rate differences. Results indicated a negative, but non-significant association between fairness perceptions and selection rate differences *(r = -0.17, p = 0.63)*. A stronger negative, but non-significant association was found between fairness perceptions and false negative rate differences *(r = -0.48, p = 0.16)*. Figure 6 shows the correlation plots.

However, as a result of asking for fairness perceptions by presenting the trade-off between two fairness criteria, these plots exhibited a curvilinear relationship, as opposed to a straight line, indicating an interplay between the two criteria. To illustrate, the algorithms adhering to demographic parity (the

Figure 6: Relationship between fairness perceptions and mathematical fairness criteria. The upper plot shows the relationship between the average perceived fairness scores and the false negative rate differences between gender groups *(r = -0.48, p = 0.16)*. The lower plot shows the relationship between the average perceived fairness scores and the selection rate differences between gender groups *(r = -0.17, p = 0.63)*.



algorithms with a small selection rate difference), were not adhering to equality of opportunity (these algorithms had a large false negative rate difference), as reported in Table 14. This could have caused the concave shape of the relation between the average perceived fairness scores and the differences in selection rates, as shown in Figure 6.

Therefore, subsequently, I performed a regression analysis to assess the effect of the selection rate differences and false negative rate differences on participants' fairness perceptions. Since participants' fairness perceptions, the dependent variables, were measured on a Likert scale, I opted for ordinal regression.

Three distinct regression models were created using the 'ordinal' package in R [69]. The first model was the simplest model: it included the perceived fairness score as the dependent variable, and the selection rate differences and false negative rate differences as the independent variables. The second model was an extension of the first model: besides the selection rate differences and false negative rate differences, it also included the groups participants belong to as independent variables. The third model, a mixed-effects model, moreover included a random effects term for participants, to control for the fact that each participant was asked to evaluate five different algorithms. I subsequently conducted an ANOVA test on the three models to find out which model performed best. This test reported that the third model had a significantly better fit than the first and second models, as evidenced by its lower AIC and higher log-likelihood value compared to the other two models. Table 19 reports the results of this ANOVA test.

Table 20 presents the results of the best regression model (model 3). This model showed a significant

Table 19: Results of ANOVA test to compare the three different ordinal regression models. Model 1 only included the selection rate differences and false negative rate differences as independent variables. Model 2 also included the group participants belonged to as independent variables. In model 3, moreover, a random effects term for participants was added. Results indicate that model 3 had the best fit.

Regression model	**AIC**	**Log-likelihood**	χ^2	**P-value**
Model 1	3943.0	-1963.5	53.714	< 0.001
Model 2	3892.4	-1936.2	54.618	< 0.001
Model 3	3687.2	-1832.6	207.128	< 0.001

negative association between the selection rate differences and perceived fairness scores *(B=-5.609, SE = 0.723, $p < 0.001$)*, indicating that as the selection rate difference increased, participants were more likely to assign a lower perceived fairness score. There was a significant negative association between false negative rate differences and perceived fairness scores *(B=-5.202, SE = 0.621, $p < 0.001$)*, indicating that as the false negative rate difference increased, participants were more likely to assign a lower perceived fairness score.

Moreover, there was a significant positive association between groups 2 and 3, and perceived fairness scores (*(B=1.046, SE = 0.280, $p < 0.001$)* and *(B=1.212, SE = 0.281, $p < 0.001$)*, respectively). This indicates that being part of either of these two groups, which received information about both the distributive and procedural fairness of the algorithms, had a positive contribution to fairness perceptions. As shown by its highest coefficient, this effect was the biggest in group 3, in which the sensitive feature *gender* was not used as a main attribute. These findings align with the results of RQ1, reported in Figure 4, in which I find that fairness perceptions are highest when given information about both the distributive and procedural fairness of the algorithms, especially when gender is not included as a main feature in the algorithm.

In Appendix A.6, the full results of all three ordinal regression models are reported.

Table 20: Ordinal mixed effects regression model with a random effects term for participants. The model allows the prediction of the perceived fairness scores (on a 7-point Likert scale) from selection rate differences and false negative rate differences between groups. Standard errors are reported in parentheses.

Coefficients	**Estimate (S.E.)**	**P-value**
Selection rate difference	-5.609 (0.723)	*<0.001*
False negative rate difference	-5.202 (0.621)	*<0.001*
Group 2	1.046 (0.280)	*<0.001*
Group 3	1.212 (0.281)	*<0.001*
Random effects	**Variance (S.E.)**	
Participant	2.21 (1.487)	

5.2 Qualitative Analysis

To gain additional insights into the findings of my quantitative analysis, I qualitatively analyzed participants' rationales behind their fairness ratings. This was done by openly coding their responses to the open-ended question of which factors they considered most important in determining the fairness of the algorithms. Although each participant provided an explanation, I encountered a variety of response lengths: responses varied in length between 1 word and 59 words, with a mean of 12 words and a median of 9 words. By first identifying first-order codes out of these responses, grouping these into second-order codes, and finally dividing these into two third-order concepts, I systematically classified the responses. Figure 7 gives an overview of these categories and provides, per category, an indicative quote. I will now discuss some of the responses falling under the two third-order concepts I identified: distributive fairness and procedural fairness.

5.2.1 Distributive fairness

While I encountered a variety of answers, the biggest proportion of explanations (n=173, 77%) could be attributed to the outcome of the algorithms, relating to the concept of distributive fairness. This was as expected, as only two out of three groups received a feature importance graph, and all three groups received information about distributive fairness. However, interestingly, I observed that across all three groups, the majority of participants focused on distributive fairness rather than procedural fairness (89% of all answers in group 1, 63% of all answers in group 2, and 79% of all answers in group 3).

More specifically, across all three groups, most participants (n=88) emphasized the importance of considering the trade-offs between the different fairness criteria shown in the graphs. For example, P45 (group 1), stated: *"I mainly looked at the proportions between genders of those qualified but not hired in comparison to the genders when hired"*.

The second most frequently mentioned category pertained to the concept of equal opportunity: a notable proportion of participants (n=51) mainly focused on false negative rates and the qualifications of candidates. This finding suggests a preference for fairness criteria that consider both the predicted and true outcome, rather than only the predicted outcome. For example, P18 (group 1) answered: *"The percentage that was qualified but not hired was the most important factor for me"*.

Nevertheless, there was also a substantial number of participants (n=34) that primarily considered the selection rates of both groups, e.g.: *"Whether the hired % of candidates were as equal as possible"* (P38, group 2). However, across all three groups, this category, associated with demographic parity, was mentioned less frequently than the category relating to equal opportunity.

5.2.2 Procedural fairness

The remaining 13% of answers (n=52) could be attributed to the decision-making process, and therefore, to the concept of procedural fairness (11% of all answers in group 1, 37% of all answers in group 2, and 21% of all answers in group 3).

The majority of these responses (n=45) were related to the features used by the algorithms. For example, in group 2, in which gender was included as a main attribute in the feature importance graph, I encountered 16 answers that explicitly criticized its usage, e.g.: *"I marked them all low as I don't see why gender would be an important factor "* (P68, group 2). Other participants mainly focused on the importance or combination of the different attributes, e.g., *"The 5 main attributes were the main thing I considered"* (P52, group 2).

Apart from the procedural fairness of using certain features, some participants did not provide reasons specific to the information shown in the graphs but criticized the use of algorithms for hiring in general (n=7). For example, P70 (group 3), wrote: *"I don't believe this kind of selection is fair in any circumstances"*, and P31 (group 1) stated: *"I don't find the process fair as I believe the candidate should have a formal interview rather than just basing the hire on grades and qualifications"*.

Figure 7: Indicative quotes, first-order quotes, second-order quotes and third-order codes for the open-ended question: *"Which factors did you consider most important in determining whether a model was fair or unfair?"*

5.3 Demographical Analysis

Additionally, I investigated whether participants' demographic characteristics had any impact on their mean perceived fairness scores. As demonstrated in the following sections, no strong effects of gender, age, education level, or experience in AI, were identified.

5.3.1 Gender

To examine the potential differences in average scores among genders, I compared the average scores for each algorithm between men and women, using a Mann-Whitney U-test (a non-parametric variant of the independent t-test). Although women gave slightly higher scores to each algorithm compared to men, the results of these tests did not reveal any significant differences, as reported in Table 21.

Table 21: Results of Mann-Whitney U-tests to compare the mean perceived fairness scores of the different algorithms among genders. Results indicate that the differences in scores between men and women are not significant at $\alpha = 0.05$.

Algorithm	Mean perceived fairness score		U	P-value
Original	Men:	2.821	5860.5	0.224
	Women:	2.946		
Demographic Parity	Men:	3.415	5747.0	0.164
	Women:	3.559		
Equality of Opportunity	Men:	3.862	6109.0	0.412
	Women:	3.869		

5.3.2 Age

To examine for potential differences in average scores among participants of different ages, I examined whether the scores between the different age groups as reported in Table 15 differed significantly. By, for each algorithm, performing a Kruskal-Wallis H test (a non-parametric variant of the ANOVA test to compare multiple groups) between the scores in these different groups, however, no significant results were found, as reported in Table 22.

Table 22: Results of Kruskal-Wallis H tests to compare the mean perceived fairness scores of the different algorithms among age groups. Results indicate that the differences in scores between age groups are not significant at $\alpha = 0.05$.

Algorithm	Mean perceived fairness score		H	P-value
Original	18-30:	2.787	1.914	0.384
	30-45:	3.050		
	45-60:	2.720		
	>60:	2.900		
Demographic Parity	18-30:	3.620	1.712	0.425
	30-45:	3.575		
	45-60:	3.360		
	>60:	3.050		
Equality of Opportunity	18-30:	3.987	0.732	0.694
	30-45:	3.775		
	45-60:	3.830		
	>60:	3.825		

5.3.3 Education level

To examine for potential differences in average scores among the different levels of education attained by participants, I analyzed whether there was a significant difference in scores between those who completed high school and those who obtained a degree beyond high school. A Mann-Whitney U-test revealed that for the algorithms adhering to demographic parity, participants who had obtained a degree beyond high school rated the algorithms as significantly fairer compared to participants who had solely completed high school. However, the observed result was not highly significant, as the p-value was only slightly lower than the significance level of $\alpha = 0.05$. No significant results were found for the other algorithms, as reported in Table 23.

Table 23: Results of Mann-Whitney U-tests to compare the mean perceived fairness scores of the different algorithms among education levels. Results in italics indicate that for the algorithms adhering to demographic parity, the differences in scores are significant at $\alpha = 0.05$.

Algorithm	Mean perceived fairness score		U	P-value
Original	High school:	2.875	6040.5	0.484
	Above high school:	2.871		
Demographic Parity	High school:	3.350	5252.5	*0.043*
	Above high school:	3.683		
Equality of Opportunity	High school:	3.929	5603.0	0.166
	Above high school:	3.762		

5.3.4 AI experience

Lastly, I investigated whether experience in computer science or Artificial Intelligence had an effect on participants' fairness perceptions. A Mann-Whitney U test between participants that either had experience or not, however, revealed that participants without experience rated the algorithms adhering to demographic parity as significantly fairer than participants with experience. For the other two algorithms, no significant differences were found, as reported in Table 24.

Table 24: Results of Mann-Whitney U-tests to compare the mean perceived fairness scores of participants with and without experience in computer science/AI. Results in italics indicate that for the algorithms adhering to demographic parity, the differences in scores are significant at $\alpha = 0.05$.

Algorithm	Mean perceived fairness score		U	P-value
Original	Experience:	2.750	4337.0	0.221
	No experience:	2.898		
Demographic Parity	Experience:	3.839	3737.5	*0.014*
	No experience:	3.389		
Equality of Opportunity	Experience:	3.946	4435.5	0.304
	No experience:	3.834		

Discussion

This section will start with a further discussion of the findings of the empirical study. Furthermore, Section 6.2, will outline the limitations of this thesis, and Section 6.3 will offer recommendations for future research.

6.1 Discussion of results

Previous studies on algorithmic fairness perceptions have primarily focused on distributive fairness, procedural fairness, or interactional fairness in isolation. However, the results of this study highlight the need to consider the interplay between these different fairness components in research into fair AI.

By considering the importance of different features used by a model as a key aspect of procedural fairness, the main finding of this study is that participants who receive information about both the distributive and procedural fairness of an algorithm, perceive it as fairer, than participants who only receive information about the distributive fairness of an algorithm (Section 5.1.1). Surprisingly, even when gender, a sensitive attribute, is included as a primary attribute in the algorithms, this effect is still observed, despite a substantial number of participants citing it as unfair in the open-ended question (Section 5.2.2).

These findings underscore the potential consequences of adopting the *distributiveness assumption* as described by Dolata et al. [55], as I show that solely representing algorithmic fairness as an outcome distribution issue can lead to lower perceptions of fairness. The results of this study suggest that providing more information about the workings of an algorithm can enhance fairness perceptions. This is consistent with the results of Dodge et al. [26] and Angerschmid et al. [66], who find that feature importance-based explanations have a positive impact on algorithmic fairness perceptions.

Furthermore, this work provides empirical insights into how mathematical fairness criteria are related to human algorithmic fairness perceptions. This is done by measuring and comparing participants' fairness perceptions of recruitment algorithms adhering to either demographic parity or equality of opportunity, and by measuring the correlation between participants' fairness perceptions and algorithmic selection rate differences and false negative rate differences, respectively.

Interestingly, a significant preference for equality of opportunity over demographic parity is found when given information about both the distributive and procedural fairness of the algorithms (Section 5.1.2). These findings are affirmed in the qualitative analysis, in which it is noted that a larger proportion of participants assigns greater importance to false negative rates when forming their fairness judgments, as opposed to (equal) selection rates among genders (Section 5.2.1).

These results are in contrast with the preference for demographic parity found by Srivastava et al. [22]. As they focus on a medical risk prediction and criminal risk prediction setting, rather than hiring, these varying contexts could be the reason behind these contrasting findings. It is, however, also plausible that these disparate results can be explained by the varying methods of visualizing fairness issues. Where Srivastava et al. [22] represent their algorithms by showing the individual outcomes of ten decision sub-

jects, this study reports the trade-offs between two fairness criteria. Moreover, while all participants in the study of Srivastava et al. [22] are solely provided with information about the algorithmic outcomes, relating to the concept of distributive fairness, two-thirds of the participants in this study also receive information about the procedural fairness of the algorithms.

In a study into lay people's understanding of mathematical fairness criteria, interestingly, Saha et al. [61] find that participants' comprehension of equality of opportunity is lower compared to their comprehension of demographic parity. Additionally, they observe that participants who score higher on comprehension tend to have lower fairness perceptions. In line with this reasoning, a possible explanation for the findings of this thesis is that the participants had a better understanding of the algorithms adhering to demographic parity compared to the algorithms adhering to equality of opportunity. This could have resulted in assigning a lower score to the algorithms adhering to demographic parity. Future studies could further investigate the relationship between comprehension of certain fairness criteria and algorithmic fairness perceptions.

Furthermore, by measuring the relation between fairness perceptions and the selection rate differences and false negative rate differences, no significant correlations are found (Section 5.1.3). However, it is important to acknowledge that the visualization of algorithmic outcomes as trade-offs between selection rates and false negative rates may have led to a flawed investigation of these correlations. For example, some algorithms exhibited a small difference in selection rate differences across groups (an indicator of fairness), but also a high difference in false negative rates across groups (an indicator of unfairness). However, by conducting an ordinal regression analysis, a negative effect of both selection rate differences and false negative rate differences on perceived fairness scores is found. These findings suggest that algorithms with greater disparities in selection rates or false negative rates are perceived as less fair by participants.

Lastly, this study presents a brief examination of how demographic characteristics influence perceptions of fairness (Section 5.3). In line with the results of Wang et al. [19] and Grgic-Hlaca et al. [25], I do not find any significant effects of age and gender on participants' fairness perceptions. However, this analysis reveals a positive effect of education level on fairness perceptions of algorithms adhering to demographic parity. This result is in contrast with that of Van Berkel et al. [60], who find a negative correlation between education level and perceived algorithmic fairness. Moreover, this analysis reveals a negative effect of experience in computer science or AI on fairness perceptions of algorithms adhering to demographic parity. This finding is in contrast with the results of Wang et al. [19] and Pierson [24], who both conclude that computer literacy increases perceived algorithmic fairness. Since this thesis only identifies demographic effects in algorithms that adhere to demographic parity, and not in other algorithms, I however refrain from making any final conclusions based on these findings.

6.2 Limitations

This study has several limitations. First, I conducted the study with crowdworkers. Although they were pre-selected on having obtained at least a high-school diploma, the possibility of some participants not understanding or being able to correctly interpret the trade-offs being shown can not completely be ruled out. I tried to keep the visualizations as straightforward as possible by showing bar graphs but acknowledge the possible difficulty of the task. As the results were consistent amongst groups, I however believe the results correctly reflect the intuitions of the participants.

A second limitation of this study relates to the features used by the models. As the data set did not indicate what kind of companies it considered, some participants mentioned they did not fully understand the particular selection of the top five most important attributes. Moreover, since I used postprocessing bias mitigation, the feature importance graph stayed the same across all algorithms, which could possibly have caused some confusion. I did this, however, to ensure the validity of studying the differences between groups. Future research could investigate the effect of different levels of feature

importance on participants' fairness perceptions.

The study has a third important limitation, regarding the examination of the correlations between fairness perceptions and mathematical fairness criteria. Firstly, there was a limited amount of selection rate differences and false negative rate differences available to correlate with participants' perceptions of fairness. Additionally, the presentation of the graphs as a trade-off led to an interaction between both criteria. In retrospect, these factors should have been taken into account while designing the experimental setup.

A final limitation relates to the participants' demographics. While the study had an even distribution of male and female participants, the vast majority of participants were White. Future research should aim to expand the representation of racial groups, to mitigate the risk of developing a one-sided and potentially biased understanding of perceived algorithmic fairness.

6.3 Future Work

This study provides important directions for further research. In particular, the results emphasize that understanding algorithmic fairness perceptions requires careful consideration of both visualization and contextual factors. Suggestions for future work, therefore, include:

- Exploring the effect of presenting various visualizations, and offering additional context about the decision-making process, on participants' algorithmic fairness evaluations. Van Berkel et al. [60], for example, take a useful start in this direction, by evaluating the effect of scatterplot and text-based visualizations of algorithmic outcomes on fairness perceptions.

- Investigating participants' preferred mathematical fairness criteria in multiple contexts, besides algorithmic hiring. For instance, a future study could categorize various contexts based on the risk-oriented approach of the AI act, which categorizes AI systems into 4 levels: unacceptable, high, minimal, or low risk [35]. Such a study could then examine whether participants' preferences for certain fairness criteria in different contexts vary based on these different levels of risk.

- Performing a broader investigation of participants' preferred mathematical fairness criteria, by considering more fairness criteria, besides the two I chose to include in this study. For example, a next study could, like Harrison et al. [23], also incorporate false positive rates. Another option could be to adopt a *human-in-the-loop approach*, where participants are presented with several algorithmic outcomes and are given the opportunity to manually adjust these outcomes. The next step would then be to evaluate whether they preferentially adjust outcomes that conform to one particular fairness criterion over outcomes that conform to another criterion.

- Studying whether participants' fairness perceptions are affected by receiving additional information about an algorithm, by conducting a within-subjects study, as opposed to a between-subjects study. For example, one approach could involve presenting participants with information about the distributive fairness of an algorithm, followed by information about its procedural fairness. By asking for their fairness perceptions at these two points in time, it could be investigated whether providing information about procedural fairness *alters* fairness perceptions.

- Enhancing the correlational analysis between mathematical fairness criteria and human fairness perceptions performed in this study, by increasing the number of data points to correlate with fairness perceptions. A possible approach to achieve this could be to offer participants a wider range of algorithms, or to present different algorithms to different groups of participants.

Conclusion

In this study, I approach the topic of perceived algorithmic fairness through the lens of organizational justice theory, focusing specifically on algorithmic hiring as a case study. This section will succinctly present the final answers to the main research questions asked in this study.

RQ1: *How do human fairness perceptions of a recruitment algorithm differ when only given information about the distributive fairness of the algorithm, compared to when given information about both the procedural fairness and the distributive fairness of the algorithm?*

By grouping participants according to the type of information they receive about the recruitment algorithms, I find that participants who are only informed about the distributive fairness of the algorithms perceive them as less fair than those who are informed about both the distributive and procedural fairness of the algorithms. Interestingly, this both holds true when the sensitive attribute gender is included as a primary feature in the algorithms, and when it is not. Therefore, I conclude that providing information about procedural fairness enhances perceptions of algorithmic fairness, even when the process of decision-making can be considered unfair.

RQ2: *How do human fairness perceptions of a recruitment algorithm differ depending on whether it adheres to demographic parity or equality of opportunity?*

Across all participant groups, I observe that algorithms that adhere to equality of opportunity receive a higher average perceived fairness score than algorithms that adhere to demographic parity. This difference is significant in the groups that are provided with information on both the distributive and procedural fairness of the algorithms. In my qualitative analysis, I further notice that a larger proportion of participants is concerned with false negative rates rather than selection rates. Based on these findings, I conclude that in the context of algorithmic hiring, the fairness criterion of equality of opportunity is preferred over the criterion of demographic parity.

RQ3: *To what extent are the disparate impact score and false negative rate differences between groups of a recruitment prediction algorithm related to human fairness perceptions of it?*

By examining the correlation between the fairness perceptions of participants and, respectively, the selection rate differences and false negative rates of various recruitment algorithms, I find a negative, but non-significant association between fairness perceptions and selection rate differences, and a negative, but non-significant association between fairness perceptions and false negative rate differences. By conducting an ordinal regression analysis, I confirm these negative effects of both selection rate differences and false negative rate differences on perceived fairness scores. However, to further investigate these correlations in a more isolated manner, I suggest conducting future research that does not present algorithmic outcomes as a trade-off between two fairness criteria.

In conclusion, the results of this study highlight the interplay between the different components of algorithmic fairness, and provide an insight into the relationship between mathematical algorithmic fairness and perceived algorithmic fairness. By performing an empirical study among crowdworkers, I

add to the growing body of literature on public perceptions of algorithmic fairness and provide important directions for future research.

Bibliography

[1] C. Starke, J. Baleis, B. Keller, and F. Marcinkowski, "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature," *Big Data & Society*, vol. 9, no. 2, p. 20539517221115189, 2022.

[2] N. Kordzadeh and M. Ghasemaghaei, "Algorithmic bias: review, synthesis, and future research directions," *European Journal of Information Systems*, vol. 31, no. 3, pp. 388–409, 2022.

[3] K. Makhlouf, S. Zhioua, and C. Palamidessi, "Machine learning fairness notions: Bridging the gap with real-world applications," *Information Processing & Management*, vol. 58, no. 5, p. 102642, 2021.

[4] K. A. Houser, "Can ai solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making," *Stan. Tech. L. Rev.*, vol. 22, p. 290, 2019.

[5] L. Morse, M. H. M. Teodorescu, Y. Awwad, and G. C. Kane, "Do the ends justify the means? variation in the distributive and procedural fairness of machine learning algorithms," *Journal of Business Ethics*, pp. 1–13, 2021.

[6] M. K. Lee, "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Big Data & Society*, vol. 5, no. 1, p. 2053951718756684, 2018.

[7] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," in *Ethics of Data and Analytics*. Auerbach Publications, 2016, pp. 254–264.

[8] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," Oct 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[9] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[10] "Ethics guidelines for trustworthy ai." [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[11] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[12] B. Richardson and J. E. Gilbert, "A framework for fairness: A systematic review of existing fair ai solutions," *arXiv preprint arXiv:2112.05700*, 2021.

[13] T. Mahoney, K. Varshney, and M. Hind, *AI Fairness*. O'Reilly Media, Incorporated, 2020.

[14] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

[15] A. N. Carey and X. Wu, "The statistical fairness field guide: perspectives from social and formal sciences," *AI and Ethics*, pp. 1–23, 2022.

[16] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions," in *Proceedings of the 2018 Chi conference on human factors in computing systems*, 2018, pp. 1–14.

[17] S. T. Dasch, V. Rice, V. R. Lakshminarayanan, T. A. Togun, C. M. Boykin, and S. M. Brown, "Opportunities for a more interdisciplinary approach to perceptions of fairness in machine learning," in *NeurIPS 2020 Workshop: ML Retrospectives, Surveys Meta-Analyses (ML-RSA)*, 2020.

[18] N. Helberger, T. Araujo, and C. H. de Vreese, "Who is the fairest of them all? public attitudes and expectations regarding automated decision-making," *Computer Law & Security Review*, vol. 39, p. 105456, 2020.

[19] R. Wang, F. M. Harper, and H. Zhu, "Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[20] A. Wang, "Procedural justice and risk-assessment algorithms," *Available at SSRN 3170136*, 2018.

[21] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. De Vreese, "In ai we trust? perceptions about automated decision-making by artificial intelligence," *AI & SOCIETY*, vol. 35, no. 3, pp. 611–623, 2020.

[22] M. Srivastava, H. Heidari, and A. Krause, "Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2459–2468.

[23] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect machine learning models," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 392–402.

[24] E. Pierson, "Demographics and discussion influence views on algorithmic fairness," *arXiv preprint arXiv:1712.09124*, 2017.

[25] N. Grgić-Hlača, A. Weller, and E. M. Redmiles, "Dimensions of diversity in human perceptions of algorithmic fairness," *arXiv preprint arXiv:2005.00808*, 2020.

[26] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, "Explaining models: an empirical study of how explanations impact fairness judgment," in *Proceedings of the 24th international conference on intelligent user interfaces*, 2019, pp. 275–285.

[27] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 903–912.

[28] M. K. Lee, A. Jain, H. J. Cha, S. Ojha, and D. Kusbit, "Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–26, 2019.

[29] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller, "Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[30] J. Greenberg, "A taxonomy of organizational justice theories," *Academy of Management review*, vol. 12, no. 1, pp. 9–22, 1987.

[31] J. A. Colquitt, "On the dimensionality of organizational justice: a construct validation of a measure." *Journal of applied psychology*, vol. 86, no. 3, p. 386, 2001.

[32] L. Li, T. Lassiter, J. Oh, and M. K. Lee, "Algorithmic hiring in practice: Recruiter and hr professional's perspectives on ai use in hiring," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 166–176.

[33] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 469–481.

[34] C. Schumann, J. Foster, N. Mattei, and J. Dickerson, "We need fairness and explainability in algorithmic hiring," in *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2020.

[35] E. Commission. (2023, Mar.) Regulatory framework proposal on artificial intelligence. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[36] S. C. Geyik, S. Ambler, and K. 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*, 2019, pp. 2221–2231.

[37] A. Köchling and M. C. 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*, vol. 13, no. 3, pp. 795–848, 2020.

[38] N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1083–1094, 2022.

[39] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[40] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on information systems (TOIS)*, vol. 14, no. 3, pp. 330–347, 1996.

[41] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[42] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[43] "About deus." [Online]. Available: https://www.deus.ai/about

[44] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and access in algorithms, mechanisms, and optimization*, 2021, pp. 1–9.

[45] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[46] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, "A clarification of the nuances in the fairness metrics landscape," *Scientific Reports*, vol. 12, no. 1, pp. 1–21, 2022.

[47] P. Gajane and M. Pechenizkiy, "On formalizing fairness in prediction with machine learning," *arXiv preprint arXiv:1710.03184*, 2017.

[48] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[49] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[50] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.

[51] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[52] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[53] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[54] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–16.

[55] M. Dolata, S. Feuerriegel, and G. Schwabe, "A sociotechnical view of algorithmic fairness," *Information Systems Journal*, vol. 32, no. 4, pp. 754–818, 2022.

[56] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 56–65, 2019.

[57] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.

[58] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in ai," *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[59] J. A. Adebayo *et al.*, "Fairml: Toolbox for diagnosing bias in predictive modeling," Ph.D. dissertation, Massachusetts Institute of Technology, 2016.

[60] N. Van Berkel, J. Goncalves, D. Russo, S. Hosio, and M. B. Skov, "Effect of information presentation on fairness perceptions of machine learning predictors," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.

[61] D. Saha, C. Schumann, D. Mcelfresh, J. Dickerson, M. Mazurek, and M. Tschantz, "Measuring non-expert comprehension of machine learning fairness metrics," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8377–8387.

[62] L. P. Robert, C. Pierce, L. Marquis, S. Kim, and R. Alahmad, "Designing fair ai for managing employees in organizations: a review, critique, and design agenda," *Human–Computer Interaction*, vol. 35, no. 5-6, pp. 545–575, 2020.

[63] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu, "How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 99–106.

[64] K. Van den Bos, H. A. Wilke, and E. A. Lind, "When do we need procedural fairness? the role of trust in authority." *Journal of Personality and social Psychology*, vol. 75, no. 6, p. 1449, 1998.

[65] G. S. Leventhal, "What should be done with equity theory?" in *Social exchange*. Springer, 1980, pp. 27–55.

[66] A. Angerschmid, J. Zhou, K. Theuermann, F. Chen, and A. Holzinger, "Fairness and explanation in ai-informed decision making," *Machine Learning and Knowledge Extraction*, vol. 4, no. 2, pp. 556–579, 2022.

[67] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

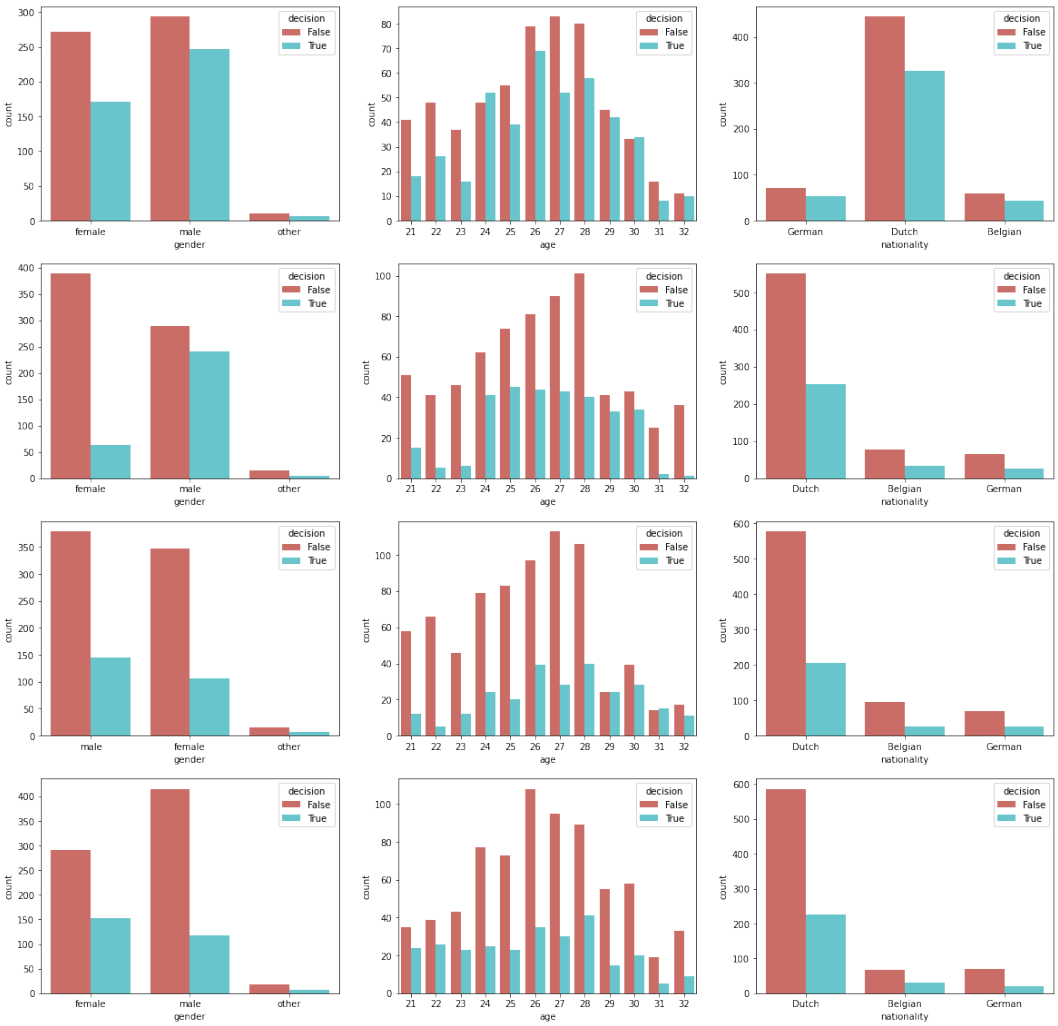
[68] Microsoft Corporation, "Microsoft excel." [Online]. Available: https://office.microsoft.com/excel

[69] R. H. B. Christensen, "Cumulative link models for ordinal regression with the r package ordinal," *Submitted in J. Stat. Software*, vol. 35, 2018.

Appendix

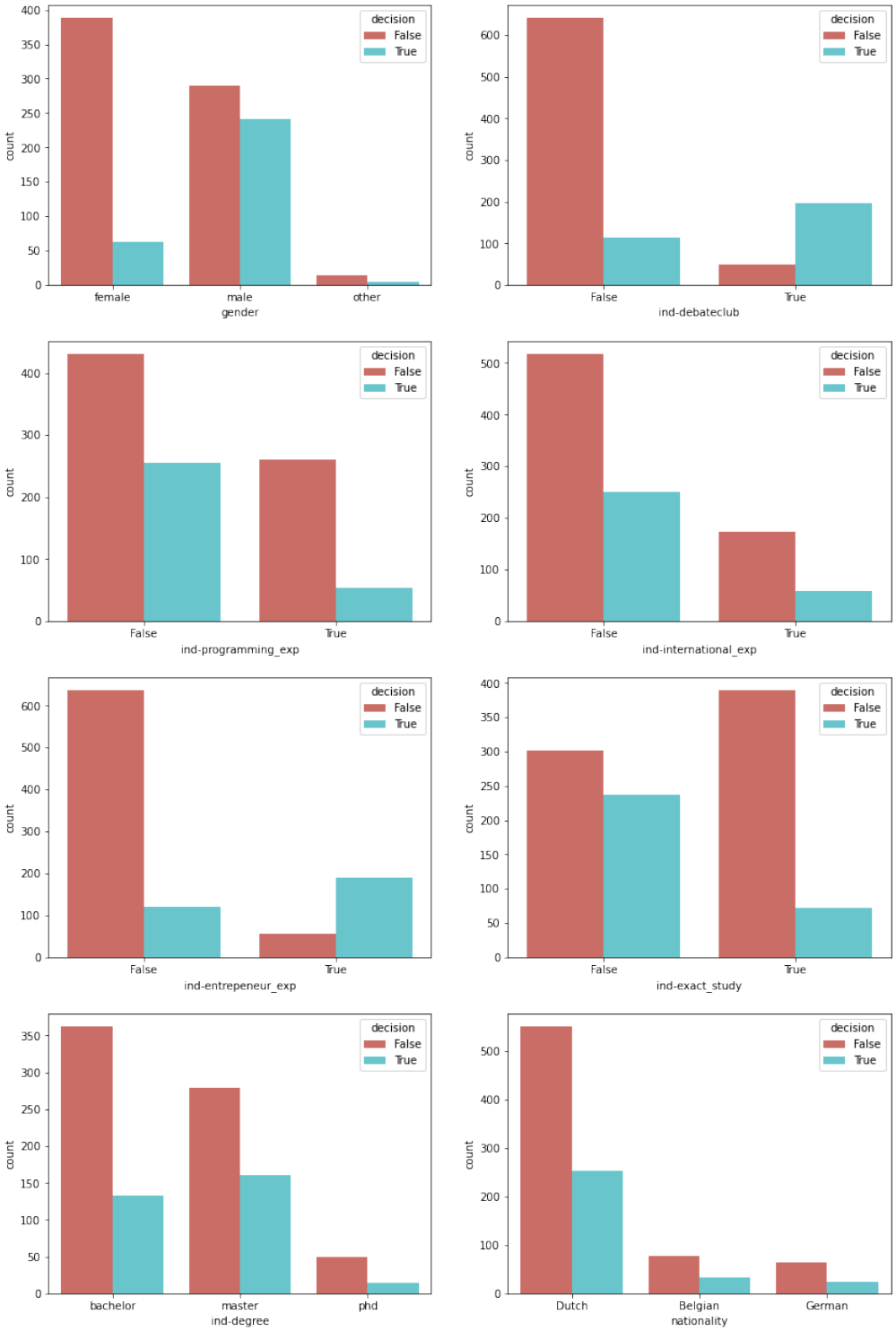
A.1 Count plots of 4 data subsets

Figure 8: Count plots showing the distribution of the target attribute (being hired to the company or not) for each of the four different companies. Each row represents one company. Note that the gender bias is largest in the company represented in the second row: here, the proportion of hired male candidates is visibly larger compared to the proportion of hired female candidates.



A.2 Count plots of final data

Figure 9: Count plots of final data subset showing the distribution of the target attribute (being hired to the company or not) among the different categorical features.



A.3 Feature importance of all features used in recruitment prediction model

Figure 10: Feature importance of all features used in the recruitment prediction model. On the left side of the table, the importance buckets of the features are reported. The six features with the biggest positive influence on receiving the favorable outcome, i.e., the six features with the highest coefficients, were used in the user experiment. This concerned the six features belonging to the three highest feature importance buckets.

Importance	Coefficient	Feature
(0.647, 9.672]	9.672093	ind-languages
(0.647, 9.672]	0.814771	ind-degree_master
(0.407, 0.647]	0.630091	ind-university_grade
(0.407, 0.647]	0.496310	ind-debateclub_True
(0.261, 0.407]	0.386807	ind-exact_study_False
(0.261, 0.407]	0.337461	gender_male
(0.215, 0.261]	0.232922	nationality_German
(0.215, 0.261]	0.217182	ind-entrepeneur_exp_True
(0.126, 0.215]	0.213152	ind-programming_exp_False
(0.126, 0.215]	0.162305	ind-international_exp_False
(-0.0963, 0.126]	0.096124	nationality_Belgian
(-0.0963, 0.126]	-0.040566	gender_other
(-0.215, -0.0963]	-0.163100	ind-international_exp_True
(-0.215, -0.0963]	-0.213946	ind-programming_exp_True
(-0.306, -0.215]	-0.217977	ind-entrepeneur_exp_False
(-0.306, -0.215]	-0.297690	gender_female
(-0.359, -0.306]	-0.329840	nationality_Dutch
(-0.359, -0.306]	-0.352961	ind-degree_phd
(-0.466, -0.359]	-0.387601	ind-exact_study_True
(-0.466, -0.359]	-0.462605	ind-degree_bachelor
(-1.13, -0.466]	-0.497104	ind-debateclub_False
(-1.13, -0.466]	-1.128784	age

A.4 Introductory text Qualtrics survey

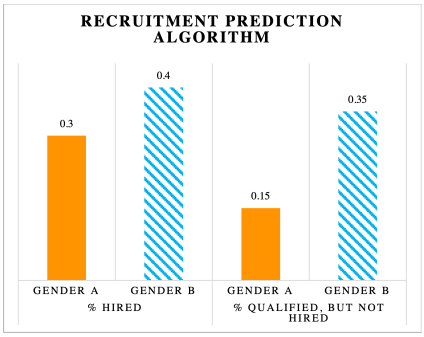
The following text was shown to the participants:

In this survey, you will be asked to answer questions about a hypothetical recruitment algorithm. Please start by carefully reading the following scenario:

A consulting company uses a computer algorithm to assist in its recruitment process. Based on different attributes of new job candidates, such as university grades, this algorithm automatically decides whether the company should hire or reject a candidate. The algorithm makes these decisions based on data about earlier job applicants that the company has collected over time.

In this survey, you will see several computer algorithms. These algorithms will be visualized as graphs showing the hiring outcomes for 2 groups: gender A and gender B. [You will also get to see an overview of the 5 attributes that the algorithm considers most important in making its decisions. Note that these attributes are the same in all different algorithms.][1] Below is an example of how the graphs showing the hiring outcomes will look.

On the left side of the graph, you can see the total percentage of candidates that are hired: this al-



gorithm hires 30% of all candidates of gender A and 40% of all candidates of gender B. On the right side of the graph, you can see the percentage of qualified candidates that are not hired: this algorithm rejects 15% of all qualified candidates of gender A and 35% of all qualified candidates of gender B. Note that these percentages do not have to add up to 100%!

In this survey, you will be asked to rate the fairness of these algorithms, focusing on how the algorithm treats the two groups differently. Fairness is broadly defined as: "the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits".

[1] These two sentences were only shown in groups 2 and 3, which also received information about the procedural fairness of the algorithms

A.5 Full survey data of 3 randomly selected participants

Table 25: Full survey data of 3 randomly selected participants. Participant 1's emphasis in making fairness judgments was on demographic parity rather than equality of opportunity, as indicated by both his perceived fairness scores and his response to the open-ended question. In contrast, Participant 2 prioritized equality of opportunity, while Participant 3's primary concern was the use of the sensitive feature gender.

Question	Participant 1	Participant 2	Participant 3
Original algorithm	2	2	1
Demographic parity - mitigated	5	2	1
Demographic parity	6	3	1
Equality of opportunity - mitigated	3	6	1
Equality of opportunity	3	6	1
In the previous questions, which factors did you consider most important in determining whether a model was fair or unfair?	*" How equal the hiring rate was"*	*"The percentage of genders that were qualified but not hired"*	*"Gender, why is that even a consideration?"*
Do you have any experience with computer science and/or artificial intelligence?	No	No	No
To which gender do you most identify?	Male	Female	Male
What is the highest level of education you have obtained?	Community college	High school diploma	Community college

A.6 Ordinal regression models

Below, the full ordinal regression models created for RQ3 are reported.

Figure 11: First ordinal regression model. The dependent variable is the perceived fairness score (PS). The independent variables are the selection rate differences (SR) and false negative rate differences (FNR).

```
formula: as.factor(PS) ~ SR + FNR
data:    all_data

 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  1125 -1963.49 3942.98 5(0)  5.02e-07 1.4e+03

Coefficients:
    Estimate Std. Error z value Pr(>|z|)
SR   -4.5749     0.6874  -6.656 2.82e-11 ***
FNR  -4.2513     0.5919  -7.182 6.86e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
1|2  -4.0350     0.2713 -14.872
2|3  -2.7227     0.2545 -10.700
3|4  -1.4035     0.2443  -5.744
4|5  -0.9247     0.2428  -3.809
5|6   0.3132     0.2465   1.271
6|7   2.0572     0.2986   6.889
```

Figure 12: Second ordinal regression model. The dependent variable is the perceived fairness score (PS). The independent variables are the selection rate differences (SR), false negative rate differences (FNR) and groups participants belonged to.

```
formula: as.factor(PS) ~ SR + FNR + GROUP_B + GROUP_C
data:    all_data

 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  1125 -1936.18 3892.37 5(0)  5.17e-07 1.3e+03

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
SR       -4.6816     0.6874  -6.810 9.75e-12 ***
FNR      -4.3190     0.5913  -7.304 2.80e-13 ***
GROUP_B   0.7548     0.1312   5.755 8.67e-09 ***
GROUP_C   0.9110     0.1328   6.860 6.88e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
1|2  -3.5804     0.2770 -12.924
2|3  -2.2355     0.2617  -8.541
3|4  -0.8671     0.2545  -3.407
4|5  -0.3733     0.2537  -1.471
5|6   0.8951     0.2586   3.462
6|7   2.6556     0.3091   8.590
```

Figure 13: Third ordinal regression model. The dependent variable is the perceived fairness score (PS). The independent variables are the selection rate differences (SR), false negative rate differences (FNR) and groups participants belonged to. This model also includes a random effects term for each participant.

```
Cumulative Link Mixed Model fitted with the Laplace approximation

formula: as.factor(PS) ~ SR + FNR + GROUP_B + GROUP_C + (1 | Participant)
data:    all_data

 link  threshold nobs logLik   AIC     niter     max.grad cond.H
 logit flexible  1125 -1832.62 3687.24 964(6406) 5.32e-04 9.2e+02

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 2.21     1.487
Number of groups:  Participant 225

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
SR       -5.6090     0.7229  -7.759 8.58e-15 ***
FNR      -5.2023     0.6206  -8.382  < 2e-16 ***
GROUP_B   1.0405     0.2802   3.714 0.000204 ***
GROUP_C   1.2115     0.2812   4.308 1.65e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
1|2  -4.5063     0.3500 -12.874
2|3  -2.6676     0.3264  -8.173
3|4  -0.8829     0.3172  -2.783
4|5  -0.2402     0.3166  -0.759
5|6   1.3892     0.3233   4.297
6|7   3.5013     0.3770   9.289
```