"It's the most fair thing to do, but it doesn't make any sense": Perceptions of mathematical fairness notions by hiring professionals

Priya Sarkar

TUDelft

“It’s the most fair thing to do, but it doesn’t make any sense”: Perceptions of mathematical fairness notions by hiring professionals

by

Priya Sarkar
4820754

to obtain the degree of Master of Science
in Computer Science
at the Delft Univeristy of Technology,
to be defended publicly on Friday, September 23, 2022 at 13:30 CEST

Program track:	Data Science & Technology	
Thesis committee:	Dr.ir. C.C.S Liem	TU Delft (Supervisor and Chair)
	Prof.dr.ir. I.R. van de Poel	TU Delft
	Dr. L. Cavalcante Siebert	TU Delft

Preface

As I come to the end of my Master study, I would like to thank all the people who made this journey possible in their way. I would like to express my profound gratitude to my thesis supervisor, Dr. Cynthia Liem for guiding and supporting me through the toughness and thins of the thesis while giving me enough freedom to explore and ideate. I thank her for all the high-level questions and perspectives that made me think about every decision and choice made in the thesis, making the discussions fulfilling. I clearly remember how after every discussion, I felt better and more confident about pursuing the topic. I admire her work and desire to make positive changes in society. She is an inspiring role model for me now. I would like to thank Han-yin Huang, Elsa Rodriguez, Fieke Miedema, and numerous PhDs and students at Computer Science for their valuable feedback and discussions throughout my thesis. This research would not have been possible without the interviewees. I thank all the participants for taking the time to participate in the study and providing much-needed and valuable insights. I would like to thank Prof.dr.ir. I.R. van de Poel and Dr. L. Cavalcante Siebert for accepting to be part of the thesis committee.

I am grateful to my parents for all their love, care, and trust throughout my journey in Delft pursuing my Master study. Their constant support was the guiding stone in this journey filled with many hurdles. I am thankful to my friends Fieke, Elsa, Dyan, Raymond, and Ingrid who lifted me when life was difficult, made me believe in my work, and went out of their way to help me when I needed it the most. The whole journey of my Master thesis and more importantly life in Delft would have been impossible without them. My friends Jane, Franco, Raquel, and the communities of Foodsharing Delft, Blue Suede Shoes, and X made the whole journey of my Master study joyful and enjoyable, making it home far away from home. I am thankful for the many Master Thesis students from the 4th and 6th floor of building 28, who made working on the thesis fun, taking lunch breaks and coffee breaks every day.

Last but not the least, I would like to thank the help and support of the many psychologists, psychiatrists, doctors, and academic counselors who saw me through my struggles with severe depression over the last few years and made me come out of it better and stronger. I can say that the journey has been memorable.

- Priya Sarkar
September 2022, Delft

Abstract

Mathematical fairness notions introduced in literature aim to make algorithmic decisions fair. However, their usage has been criticized in domains such as recidivism and lending for producing unfair decisions. Questions regarding fairness, which also have an important role in hiring are giving way to concerns about the increasing adoption of algorithmic decision-making systems in the field. However, there are no concrete studies linking mathematical notions to actual perceptions of fairness by people active in hiring and applicant selection.

We aim to explore the understanding and alignment of existing fairness notions by organizational representatives in the context of early candidate selection in hiring. Towards that, we interviewed 17 professionals from executive functions, talent acquisition, HR, I/O psychology, and diversity and inclusion operations in The Netherlands. By designing user-friendly illustrations and explanations in the context of early candidate selection in hiring, we explore their ratings and responses to six fairness notions on understandability, perception of fairness, perception of diversity, and applicability. Our qualitative investigation suggests that these fairness notions raise three concerns. One, they lack additional contexts such as a company's size or diversity goals. Two, they give rise to several ethical and practical concerns such as lack of trust in the data, disadvantaging minorities, or the selection of unqualified applicants. Lastly, they act only as a small step towards fairness in the large hiring pipeline. We conclude that a qualitative approach in collaboration between designers, practitioners, and policymakers is the key to refinement and contextualization of future technologically enabled fair hiring policies. Our participants' intrinsic motivation to engage with the topic of fairness strengthens our case.

Table of Contents

1 INTRODUCTION

Data-driven algorithmic decision making systems are increasingly being adopted in many domains, from low stakes situations such as re-routing passengers on an overbooked flight [18], to high stakes situations such as predicting incarceration or parole in criminal defense systems [4]. While in the low stakes application of re-routing passengers the algorithmic decision predicts or selects a number of passengers to re-route, in the high stakes application, the algorithmic decision predicts who will go to jail for how long. Several studies show the applicability of these data driven automated decision making systems in cancer prediction, credit scoring and lending, hiring, and university admissions. Needless to say, these data-driven automated algorithmic decision making systems can impact a person's course of medical treatment, employment, economic or educational opportunities.

The systems designed for these applications are powered by machine learning (ML) algorithms that are trained to learn from large quantities of data obtained from the domain of application. Upon learning to identify patterns, the algorithms are able to make predictions on new, unseen data. For example, if an algorithm is trained on the past employees' performance data, it can, given the data for a new employee, predict his or her performance. However, with the data-driven algorithmic decision making systems trained to identify and replicate major patterns in the data, leading to higher accuracy of the predictions made, there are risks of harmful patterns such as historical prejudice or even discrimination being replicated in the new predicted decisions [90], [23]. In the same example of predicting employee performance - if the data mainly contains records of employees with certain ethnic or educational background receiving good performance, it is highly likely that a system trained on this data will predict similar results for similar people. The historically disadvantaged people do have enough representation in the data. Moreover, the data may only reflect their representation in the negative category. For example, most of the data for recidivism prediction for ethnic minority shows them receiving stricter sentences. While the algorithms are trained on data containing harmful biases from the past, it is no longer acceptable nor legal to make discriminatory decisions today.

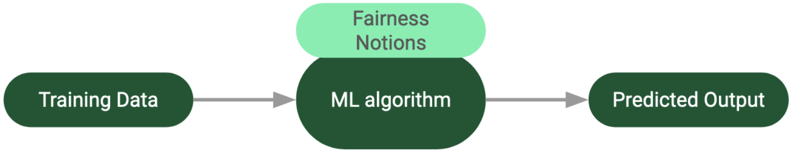


Figure 1: A simple Fair ML pipeline

In recent years, innovations to address and mitigate algorithmically driven discriminatory decisions have come in the form of mathematical fairness notions - quantitative constraints that seek to limit or check the number of negative decisions given to historically disadvantaged people, known as minority groups. A simplistic Fair ML pipeline is shown in Figure 1. The proposed mathematical notions, giving rise to the field of Fair ML [9], [33], [16], [101], aim to handle the societal disparities that – if not handled in algorithmic systems – can lead to discriminatory outcomes and is thus, unfair [9]. The societal disparities in our context originate from sensitive features that differentiate people, such as gender, ethnicity, religion, national origin, disability, age, etc.

Many fairness notions have been solely proposed based on data and sometimes, domain knowledge. In our literature survey, in Chapter 2, we identify 17 definitions, which can be classified broadly under individual or group fairness. Individual fairness notions look for similar predictions for similar individuals, whereas, group fairness treats social groups as a whole, equally [80]. For instance, one of the earliest, yet still discussed in recent literature, group fairness notions is Statistical Parity [43], which requires equal prediction rates across groups defined by their sensitive feature. Another notion, Fairness through Awareness [43], based on individual fairness, asks for the same prediction for any two individuals in the data, regardless of their sensitive feature. Multiple reasons make this field challenging; many notions cannot be satisfied simultaneously [76], [4], difficultly reconciling group and individual fairness notions [17], and no guidelines on operationalizing fairness notions [139].

Despite these challenges, fairness notions have been applied to real life situations and have subsequently received substantial criticism in domains such as recidivism and lending for producing decisions that are still discriminatory against minority groups [23]. The domain of hiring, which has not been as extensively researched as recidivism and lending, has seen increasing adoption of algorithmically driven decision making systems in the form of automated assessment of applicants [89]. However, there is no clarity on how these automated assessment systems produce non-discriminatory and bias-free decisions [111]. This brings uncertainty about the ability of mathematical interventions to solve societal biases.

An interdisciplinary approach is required to bridge the gap between designers, who think computationally and users, who think contextually about achieving fairness in a societal domain. In the hiring domain, this is relevant as organizations try to shape their policies on fair hiring - with or without algorithmic decision making systems - through their Equity, Diversity and Inclusion operations. Towards that end, the aim of this research can be summarized by the following main and sub research questions:

RQ : How do organizational representatives understand and perceive different mathematical fairness notions in the context of early candidate selection in hiring?

- RQ1 : How do organizational representatives align with different mathematical fairness notions in the context of early candidate selection in hiring?
    1. What is their conceptual understanding of the mathematical fairness notions?
    2. What do they rate the mathematical fairness notions (and their output decisions) in terms of fairness and why?
    3. What do they rate the mathematical fairness notions in terms of improving gender balance in organizations?
- RQ2 : What is their impression, perception of fairness, perception of diversity and applicability of using the different mathematical fairness notions during the early candidate selection in hiring?

To capture the understandability of the mathematical fairness notions by organizational representatives, we investigate their conceptual understanding of the notions. To capture their perception, we examine their reasoning behind four key components - their impression, perception of fairness, perception of diversity and applicability of the fairness notions. We define the scope of the research to six mathematical fairness notions; chosen to cover both group and individual notions, are relevant in recent literature and are potentially applicable in the hiring domain. The four group fairness notions are Statistical Parity [43], Equal Opportunity [59], Calibration [32] and Disparate Impact [92] and the two individual fairness notions are Fairness through awareness [43] and Counterfactual Fairness [80]. Finally, we focus on candidate selection, which is earlier in hiring pipeline compared to the final hiring decision. This choice is to prevent algorithmic systems dictating final hiring decisions and having a responsible human in the loop [125], [114].

we interviewed 17 professionals from executive functions, talent acquisition, HR, I/O psychology, and diversity and inclusion operations in The Netherlands. By designing user-friendly illustrations and explanations in the context of early candidate selection in hiring, we explore their ratings and responses to four group and two individual fairness notions. Our qualitative investigation suggests that these fairness notions raise three concerns. One, they lack additional contexts such as a company's size or diversity goals. Two, they give rise to several ethical and practical concerns such as lack of trust in the data, disadvantaging minorities, or the selection of unqualified applicants. Lastly, they act only as a small step towards fairness in the large hiring pipeline. We conclude that a qualitative approach in collaboration between designers, practitioners, and policymakers is the key to refinement and contextualization of future technologically enabled fair hiring policies. Our participants' intrinsic motivation to engage with the topic of fairness strengthens our case.

The contributions of the research is threefold. First, existing literature on human centered approaches focuses on lay people's perception of fairness [115], [121],[54], [84] where most of the studies ask participants to choose the better algorithm and respond to limited questions [121], [18], [41]. In contrast, our research focused on in-depth interviews with professionals working and committed to the field of fairness in hiring, who can reflect on the topic from multiple angles. This enabled the participants to express their goals, doubts and challenges, allowing the move towards bridging the gap between designers and practitioners. Second, most of the existing work in Fair ML covers domains such as recidivism and lending [4], [115], [54], whereas we expand the discussion to the hiring domain, where research is limited to automated assessment of job applicants [90]. Moreover, we open the discussion in the field on a domain where transparency is a challenge as commercial vendors keep the workings of their debiasing solutions confidential for reasons of propriety. Third, our example illustrations of the six mathematical fairness notions enabled ease of communication with the participant stakeholders who are not necessarily mathematically aligned. Our design based on [121], was improved over multiple iterations to create both textual and visual simplicity for approaching the study, opening pathways for collaboration between stakeholders in this field.

In this thesis, we present an extensive literature survey in Chapter 2, followed by a description of the initial domain interview that helped shape and scope the research in Chapter 3. Next, we describe the methodological set-up in Chapter 4 and present the results of the research in Chapter 5. We conclude the thesis with a discussion of the results and research in Chapter 6, which ends in the final conclusion in Chapter 7.3.

2 LITERATURE SURVEY

An extensive body of work from different scientific disciplines has broached the subject of discrimination and fairness. Recently, with the risen amount of real-word applications of algorithmic decision making, the topic of discrimination and fairness has also taken a center role in machine learning research. To understand the implications of different fairness notions on the end-users of machine learning systems for hiring, we conducted a thorough literature survey. We start by exploring the definition of discrimination, why it is wrong, the different types of discrimination and the meaning of *fairness* in section 2.1. Next, we explore the hiring pipeline, the existence of discrimination in the hiring process and the consequences of discrimination in section 2.2. There is an increasing adoption of algorithmic decision making in hiring, which can lead to discrimination. We describe this and the focus on fair hiring in section 2.3. To better explain how algorithmic hiring can lead to discrimination, we detail the concept of bias in the machine learning pipeline in section 2.5. From there, we move towards describing multiple mathematical fairness notions proposed in literature in section 2.5. These fairness notions can be extracted from data distributions, confusion matrices or evaluation scores. The notions itself can be based on similarity metrics, causality or disparate treatment and impact. To point out the different fairness for individuals, groups and sub-groups, we report on the granularity of fairness notions in section 2.6.

Fair machine learning is based on different assumptions and choices when it comes to the dataset used, the input features, the predicted outcome and the algorithm used. We describe this in section 2.7. To then combine the fairness notions in machine learning to achieve these fairness notions in Fair ML, different approaches are used. We report on these approaches in section 2.8. To achieve fairness notions in machine learning application is actually quite challenging, as many previous literature has pointed out. We summarize the main bottlenecks in section 2.9. There are some challenges that are specific to the selection of job candidates, which we show in section 2.10. To bridge the gap between research on fairness notions & Fair ML and actual end-users of these systems, various human-centered approaches have been studied. We describe those in section 2.11. Qualitative research has provided us with some recommendations for practitioners of Fair ML, we summarize those in section 2.12. Finally, we conclude what is missing in the current body of literature in our research gap in 2.13.

2.1 Discrimination and Fairness

In this section, we review the meaning of discrimination, why it is considered wrong and the meaning of fairness.

2.1.1 What is Discrimination?

Discrimination is the difference in treatment received by people from varying social groups on the basis of decisions made by authorities who can distribute either benefits or harms to members of a social group [16]. Social group can be based on, for instance - race, religion, gender, age or disability. The beneficial or harmful treatment people receive can come from major decisions such as loan application, job application, parole outcome, university admission or everyday situations such as freezing of bank account, varying car insurance premium or re-routing of passengers in over booked flights. It is discriminatory or unfair to provide different treatment to people on the basis of their membership to a social group. For example, when a male or female who are equally qualified for a job but one gets hired over the other due to the employer's partial preference of one gender accounts for discrimination in hiring. Not hiring a person due to attributes not affecting their productivity in the job is considered discrimination

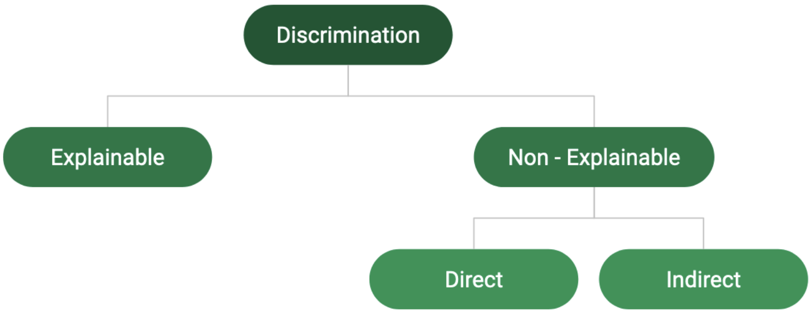


Figure 2: Types of Discrimination

[48]. Discrimination can be checked by examining how the actions performed in given situations affects people on different grounds [139]. The actions are forms of discrimination such as granting or denying a loan, the situations are the areas of discrimination such as access to employment or social benefits, and grounds of discrimination are protected characteristics such as disability or religion. Thus, it is worthwhile to understand what makes discrimination wrong and why it is problematic.

2.1.2 What makes it wrong?

This understanding can come from looking at the cause and effect of discrimination. Discrimination comes from the authorities who make decisions. Its origin can be derived from their values and belief systems, motives and intentions, implicit and explicit prejudice for or against social groups or simply put, a limited understanding of the effect of differential treatment of social groups [16] [116]. The consequential results are harmful. By failing to treat people equally implies an underlying disrespect and undermining the worth of people as human beings whether or not the differential treatment was performed intentionally or unintentionally. The effect of discrimination can be seen in the continuation of historically faulty social constructs where for instance a person may not be able to get a loan, a job, parole or university admit purely on the basis of their membership to a social group. For example, a non-native female candidate may be rejected a job interview on the basis of gender and national origin, despite meeting all the qualifications required by the employer. In this example, the employer is discriminating on factors that are not relevant to the job performance of the candidate.

2.1.3 Types of discrimination

There are two types of discrimination - explainable and non-explainable discrimination as seen in Figure 2. When discrimination between people or social groups can be explained on the basis of other factors, it is not considered illegal. To understand explainable discrimination, take for instance, the average annual income of women is lower than men in the UCI Adult dataset [7] but this can be partly explained by looking at difference in the working hours of women and men in the dataset. Women work fewer hours compared to men. Hence, the average income of women is bound to be lower than men [69]. Whereas, when discrimination between people or social groups cannot be justified on any reasonable basis, it becomes illegal. Non-explainable discrimination is further divided into direct and indirect discrimination [99], [139], which can be seen from two examples in Figure 3. It is direct discrimination when explicitly based on a sensitive factor such as religion or age. For example, if a healthcare worker is not hired by a hospital because his religious affiliation is Muslim, then the hospital is directly discriminating on the basis of religion. Whereas, discrimination is indirect when based on factors which are influenced by sensitive factors. For example, an archaeology student with parents from an immigration background has access to education only within her vicinity or economic capacity. If she is rejected from a job because of her educational background despite being qualified for the position, the employer is discriminating indirectly on the basis of national origin.

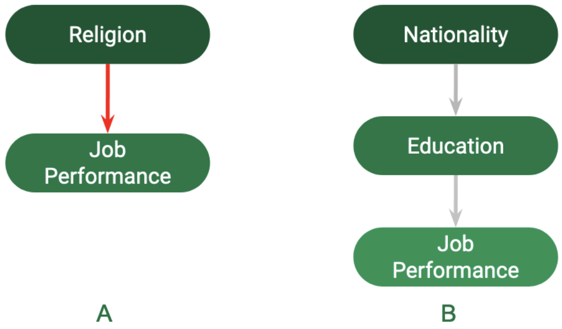


Figure 3: Examples of Discrimination. A (left) shows direct discrimination when religion, as a sensitive feature is used for predicting job performance of a job candidate, and B (right) shows indirect discrimination where nationality is the sensitive feature, which effects the level of education, which finally is used for predicting job performance.

2.1.4 Fairness

Fairness is a broad concept encompassing the philosophies behind a good life for individuals in every dimension of life. It is incorporated in multiple spheres of justice and definitions can change based on the domain and context it is applied in [16]. The same principles can manifest in different ways such as social justice, fair treatment or fair division and allocation. Narrowing it to the scope of this thesis, fairness is seen within the purview of justice - the legal and illegal decisions made by individuals or institutions using an algorithmic decision making system that hinder individuals from living a dignified life, socially or economically. Fairness in decisions given by algorithmic decision making system, also referred as machine learning (ML) or AI in mainstream literature, is important. Decisions cannot be called fair when different decisions are given for similar individuals on grounds of attributes they cannot choose such as ethnicity, disability or national origin. Discrimination on such grounds is illegal and this is the broad concept of fairness that ML systems need to incorporate. The importance of fairness in such systems can be motivated by policies and socio-economic research. Without fairness consideration, prediction based decision systems can propagate inequality and historically faulty social injustice against some social groups [9].

Decisions are fair when outcomes are not prejudiced or biased, nor show favoritism towards some individuals or groups based on their sensitive characteristics [99]. Decisions such as selecting or rejecting a person for a job is a life-changing decision that can have beneficial or harmful consequences for the person. Those decisions are discriminatory if they are based on sensitive characteristics, which can be irrelevant to the job performance [105]. When algorithms make these decisions there are multiple possibilities of discrimination and unfairness [85], [91]. Therefore, ensuring fair outcome is important. With the rise of commercial usage of AI in decision making, it has become important to incorporate measures to ensure fairness [111]. A fair machine learning model should have similar performance or outcomes across multiple social groups irrespective of the attributes such as ethnicity, religion or sex [101]. This field is referred to as Fairness in Machine Learning or Fair ML[9], [33], [16], [101].

2.2 Discrimination and Fairness in Hiring

Having reviewed the basic concepts of discrimination and fairness, in this section, we direct the same topics towards the field of hiring.

2.2.1 The Hiring pipeline

The complete hiring pipeline can be termed as Human Resource Management(HRM), which recently also inclusion various Equity, Diversity and Inclusion operations. Broadly, it can be broken down into three stages - recruitment, selection and job offering, as seen in Figure 4

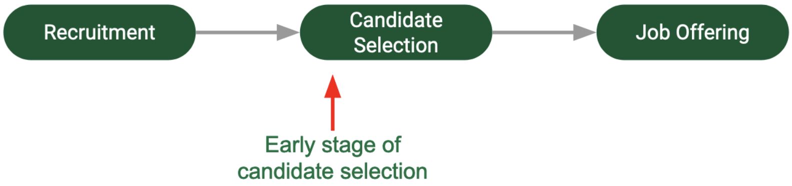


Figure 4: Out of the three stages in the hiring pipeline, early candidate selection is the focus of this research

Recruitment starts by defining the job position and advertising the vacancy through the organization's preferred mediums. The HR function termed as talent acquisition can also search for suitable candidates for the job positions, which can be performed by the employer or outsourced to recruitment agencies. These result in a number of candidate applications who can apply for the position through CV, resume, motivation letters, job application forms, etc. Next, selection stage picks job applicants from the applicant pool for the next rounds such as interviews. This stage evaluates all the applicants who have applied for the job and screens them to find a subset of candidates whom the employer determines as match for the job position. This can happen in one stage screening or multiple steps, where a combination of HR, professionals who created the job positions and employers make decisions. The selected job candidates move on to the next stage of the hiring process. In the last stage, the selected candidates go through interviews, various tests or tasks before a smaller subset gets offered the job.

The focus of this thesis is on the early selection stage which is indicated in Figure 4. It starts with a list of candidates who have applied for a job position and ends with a subset of job candidates who are deemed as potential fit for the job position. Ideally, this subset of candidates must be diverse, fairly chosen and match the requirements for the job.

2.2.2 Is there discrimination in the hiring process?

A closer look at discrimination in hiring, reveals differential treatment on the basis of ethnicity, gender, religion, disability, age and sexual orientation in a 11 year period since 2005, in multiple countries from North America, Europe, Asia, South America and Australia, with the majority of the studies conducted in the U.S and Europe [8]. The Table 1 lists the sensitive attributes according to US federal laws [83] and EU Legislation Directive [126]. To make the discussion more concrete, we present some examples of different grounds of discrimination that has surfaced in an abundance of literature. First, the probability of hiring has been shown to be influenced by sex. Men have been discriminated against when the occupational characteristics has been historically for 'female jobs' such as secretarial positions, whereas women have been similarly discriminated against in 'male jobs' such as engineering positions [112]. Moreover, findings

Sensitive Attribute	US (EEOC)	EU (Legislation Directive)
Race	Race and national origin	Race and ethnic origin
Gender	Gender and pregnancy	Gender
Religion	Religion	Religion and belief
Disability	Disability	Disability
Age	Age (above 40 years)	Age
Affiliations	Military service or affiliation	-
Economic status	Wealth	-
Genetics	Genetic information	-
Citizenship	Citizenship status	-
Other	Motherhood, marital status, sexual orientation and gender identity, political affiliation, union affiliation, physical appearance	Sexual orientation

Table 1: The sensitive attributes against which discrimination in employment is prohibited. The US federal laws (EEOC) [83] prohibits discrimination on 9 grounds and the other factors are considered illegal at least in 1 state. The European Union (EU) legislation Directive [126] prohibits discrimination on 5 grounds

show that familial status affects women more than men - married women are statistically less likely to get a callback compared to men [42]. Second, ethnicity has been historically, a big discriminatory factor at the point of hiring for multiple occupations [13], [113], [127]. Moreover, analysis of 25 year trend since 1989 shows no evidence of decline in the level of hiring discrimination against some ethnic groups in the U.S Labor market at the point of hiring, especially for African Americans [110].

2.2.3 Consequences of discrimination on social groups

Consequences of discrimination in hiring act as a negative outcome for individuals or social groups because of sensitive features they cannot control. When people with disabilities receive differential treatment, the individuals or the social groups made up of people with disabilities is affected. Therefore, a sensitive feature makes an association with the social group that can distinguish them as protected versus non-protected group, minority vs majority or advantaged versus disadvantaged group.

Social association which is out of an individual's control when used for discriminatory treatment has many harmful effects. When a person from the disadvantaged group is rejected for the job, the rejection to a job application can be detrimental [105] and have serious consequences for The consequences of rejecting a qualified candidate for a job has detrimental consequences not only for the individual is rejected for the job, the rejection to a job application can be detrimental [105] but also for the company and society regarding ethics, economic growth and procedural and distributive fairness [78].

2.3 Algorithmic decision making in Hiring

Many technical interventions have been adopted in the hiring pipeline. In this section, we discuss the types of processes that are being adopted and why, the potential discriminatory decisions they make, and finally discuss the types of discrimination technical interventions are looking to address.

2.3.1 Increasing Adoption

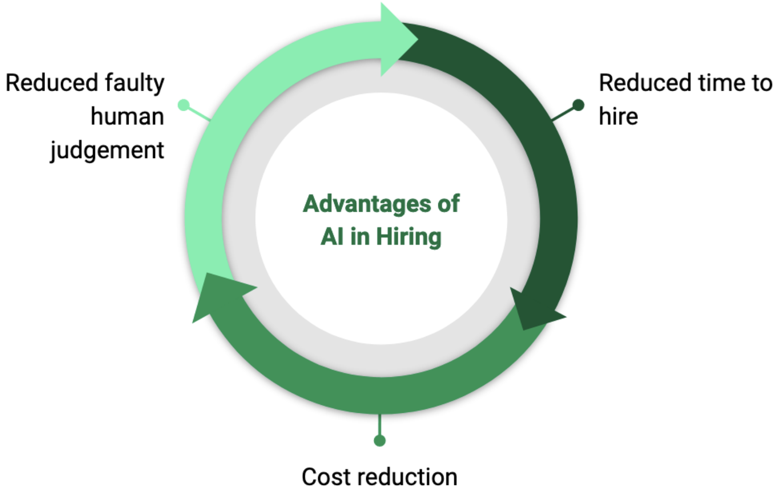


Figure 5: Reasons for increase in adoption of algorithmic decision making in hiring

Algorithms, particularly in the field of AI, are increasingly being adopted in hiring at least for an initial screening or selection of candidates [28], [111], [20]. Many arguments have been proposed in favor of this adoption, which are shown in Figure 5. First, human judgment is prone to faulty decision making based on conscious or unconscious bias which has been unresolved despite employees trained for unconscious bias [60]. ML algorithms trained for this purpose have been shown to provide decisions that can overcome faulty human judgment [77], [90]. Second, AI has improved efficiency in recruiting. There is a shift in how people are coming to be viewed as human capital who are an invaluable asset to an organization. Hence, employers look for a candidate pool with required skills as well as cultural and organizational fit. Thus, algorithms have helped in significantly reducing the time-to-hire, benefiting both the candidates and the companies [19], [90], while also achieving other organizational goals such as diversity. This efficiency in recruiting is much needed as the the number of job applications received by organizations have immensely increased, which makes it impossible for humans to process every application. For example, Walmart received on average 23,000 applicants for 600 positions for a new store [97], which can be challenging for employers to process and find the desirable candidates. Thus, achieving quality hiring based on multiple qualitative and quantitative criteria in a short period of time is impossible with only human effort. Lastly, this has also resulted in economic benefit as organizations see a reduction in cost for hiring due to the [3], [90]. Therefore, it is not surprising that automated decision making using algorithms, especially for selection or screening of candidates for jobs, is gaining more support [28], [36].

2.3.2 Discrimination from Algorithmic Hiring

However, despite the widespread use of AI in hiring and recruitment, there is a big downside that has been discovered. Machine learning algorithms are prone to providing outcomes which can be discriminatory [90]. They are data driven methods with the potential to exploit characteristics of the data it is trained on to attain the outcomes generating the best predictive model, thereby reinforcing existing human bias [111], [23]. This is referred as algorithmic discrimination and can be understood with an example. We have provided examples in previous sections where people were discriminated against in the past at the point of hire on the basis of sensitive attributes such as sex. The composition of sex in certain occupations released in 2017 by the Bureau of Labor Statistics, shows 80% of librarians were women whereas only 20% of computer programmers were women [9]. A data sample representative of this distribution of sex is not inaccurate, however, the distribution of sex over the occupation can get picked up by the machine learning pipeline and produce outcomes that reinforce the status quo. This data when used to train machine learning models comes from the past and is representative of bias and unfairness in hiring. When models are trained on such data, they also predict outcomes for new candidates which is representative of the historical bias. When the predictions or outcomes are biased towards some individuals or social groups, the algorithmic decisions are unfair or discriminatory. This implies that the stakeholders using machine learning, such as scientists, recruiters or policy makers should be aware of the impact of the decisions and ensure that ML does not propagate discrimination. In computer science literature, the sensitive factors such ethnicity, sex or religion of an individual are refereed to as "sensitive" or "protected" features. Therefore, the factors listed in Table 1 will henceforth, be refereed as sensitive features.

2.3.3 Focus of Fair Hiring

Much of the focus of fair machine learning has been on non-explainable discrimination, both direct and indirect [58], [57], [56]. This can be justified because direct and indirect discrimination has legal consequences for organizations that rely on machine learning as part of their hiring decisions. We are not implying that explainable discrimination should not be addressed. However, it can be addressed by policy level changes rather than algorithmic changes. Therefore, algorithmic decision making can work on addressing non-explainable discrimination, while leaving only explainable changes which should be handled at policy levels [69]. Our decision can be justified by examining the sources of discrimination namely, systematic and statistical [99]. Systematic discrimination arises from an organization's policies, customs or values, which is a core part of understanding explainable discrimination, making it a study beyond our scope of research. Statistical discrimination on the other hand, is caused by inferring statistical attributes to generalize group behavior, which is within the scope of our research. The inference about a group may not be correlated with fair outcome for the groups nor individuals [16]. This is particularly harmful and unfair when group characteristics are used to make general decisions about social and economic benefits that impact individuals [139].

2.4 Bias in ML Pipeline

We start the discussion of technical algorithmic decision making systems with bias. Bias, in our context refers to undesirable patterns or behavior in the system that can result in discriminatory decisions. We present the types of bias in various stages of an ML pipeline and the approaches from literature on how to target and address them.

2.4.1 Types of Bias

A machine learning pipeline consists of three stages, which can be identified as - input data, predictive modeling algorithm and outcome evaluation. These are the stages where bias can be detected and addressed, thereby consequently, lowering discrimination and moving towards a fair algorithmic model. This section outlines the areas where bias can be detected and henceforth, steps can be taken to address them. We keep the discussion specific to algorithmic decision making at the point of early candidate selection in hiring.

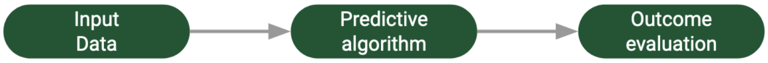


Figure 6: Three stages in the traditional machine learning platform where bias can be detected are Input Data, Predictive algorithm and Outcome Evaluation

Bias in Input Data

The patterns of discrimination in the real world gets translated into the data, which is the primary source of bias in the machine learning pipeline. This can produce data with incorrect measurements, labels, or representation of the target population. Such a dataset is influenced by many types of bias. We identify 6 types of bias in the *data* stage as follows :

- Historical Bias : When data reflects the general bias in the society, similar underlying socio-cultural patterns get embedded in it [123]. They are the real-world biases which is representative of the the world as it is [124]. For example, if the proportion of the workforce of a manufacturing company is predominantly male, a dataset of the employees from this company will reflect similar proportion.
- Representation/sampling bias : When there is a mismatch between the data sample and the world as it is, the dataset is not representative of the population on which the model will be applied to [123]. Specifically, this type of bias exists in the data when there are no data samples representing some social groups, intersections of various social groups, or disadvantaged groups receiving a favorable outcome. For example, consider a case of a financial establishment in an urban setting who have been recruiting from a top league university located in the same city. If a predictive model is trained on a dataset derived from the existing employee base, it will learn to predict that job candidates from local regions are good for hiring because it has been trained on a dataset where local job applicants were hired and applicants outside this region were either not represented or rejected as per the dataset.
- Measurement bias : The way features are measured or reported can result in measurement bias in the data sample [123]. What is measured and how it is translated into numbers or categories can be subjective and hence, produce differential measurement error, where proxy measurements differ across groups [128]. Moreover, the accuracy of measurement may vary across groups or can be an oversimplification for a complex system [124]. For example, using a "hirability score" as a measure of how likely it is that a candidate will perform well, if hired may not be psycho metrically valid but is used as a proxy measurement to determine if the job candidate should indeed be hired.
- Population bias : There is population bias in the data when the properties of the sample are different compared to the population on which the predictions will be applied to. This is similar to representation bias because of missing representation of the target population in the training data. However, this additional bias is introduced when many characteristics of population differs from the target population [104], which can be understood with an example. Consider the scenario where a machine learning model is trained to predict whom to select for entry level sales position at a company. If this model is then used to predict whom to select for a managerial level sales position or entry level marketing position, the model predictions will not be trustworthy because it was trained on a population sample with different characteristics compared to the target population.

- Self-selection Bias : This is a sub category of representation bias where individuals select themselves and there is therefore no data available on individuals who are qualified but for some substantial reason chose not to apply for the job [99]. For example, in the survey by the Bureau of Labor Statistic in 2017 [9], 90% of the nurses were female. This could imply that majority of applicants to nursing occupation were women, who selected themselves for the application. If a dataset comes from this distribution, there is no data on men who did not apply despite having the potential to train in this field, hence making such a dataset prone to self- selection bias.

- Temporal Measurement Bias : The changes in data over time attributed to evolving social definitions, behavior or population characteristics can be categorized as temporal measurement bias. For example, an article [5] presents demographic information from 1980 to 2015 from over 100 universities in America and argue that the ethnicities - "black" and "hispanic" - are still underrepresented in freshmen enrollment. However, they do not take into account the introduction of new category called "multiracial" which was introduced in 2008 [9]. Lower representation of some ethnicities can be due to the introduction of more ethnic categories. The way ethnicity is measured can evolve over time and hence, is not stable.

Whether or not, each type of data bias can be present in the final training dataset or to what extend each of the biases can be handled for fair representation is worth considering. At the point of hiring, it is important to realize that conscious choices can be made to lower or remove some data biases, which will be discussed in the next section. It is also important to be aware that the choices can produce differing outcomes from the next stages of the algorithmic model. The choices are guided by studying the importance of having a fair and representative applicant pool in the training phase. The definition of fairness, policy and organizational goals, and deliberate decisions to improve fairness in the hiring process by multiple stakeholders will help decide the biases that can be handled and its effect of partially of fully them.

Bias in Predictive algorithm

The patterns of discrimination in a skewed and biased data sample can be detected by the predictive model [16], which can produce discriminatory outcomes [9]. The models are designed to fit the input data and biases present there can be picked up by the algorithm, if not handled [33]. Many factors affect the working of the predictive model- namely, feature set used, class of the model, its functional form, the choice of model parameters, or the amount of desired interpretability [101]. The biases in the *Predictive modeling algorithm* stage are as follows:

- Omitted Variable Bias : This bias is present when variables or features containing important or relevant characteristics are not present in the model training [34]. For example, one of the features used to predict the acceptance of a job candidate is to look at previous work experience. However, excluding features such skills or volunteering experience would imply omitted variable bias, if these features are correlated to the target outcome.

- Aggregation bias : When group's behavior is not reflective of individual behavior or the group's behavior is not correlated with the expected outcome of the the decision maker, the model is prone to aggregation bias [124]. Such a bias can exist when relation between input data and target outcome is inconsistent across data or a feature value can have differing contextual meaning across subgroups [124]. This can lead to false conclusions on individuals based on entire population of a group. An algorithmic model that ignores individual differences may not be optimal where differences can exist within subgroups.

Input Data	Predictive Algorithm	Outcome Evaluation
Historical Bias	Omitted Variable Bias	Evaluation bias
Representation/sampling bias	Aggregation bias	
Measurement bias	Algorithmic Bias	
Population bias		
Self-selection Bias		
Temporal Measurement Bias		

Table 2: List of biases in the three stages of algorithmic pipeline.

- Algorithmic Bias : The bias absent from data, but added by the algorithm can lower the suitability of the autonomous system. Using a biased estimator, smoothing or regularization can modify the outcome of the algorithm [39], which is not always harmful. For instance, "biasing" the algorithm to intentionally change the outcomes to make them align with ethical goals can be helpful. However, the systems are complex and must be used with caution. Algorithmic bias has also been labeled as learning bias, where the different choices in modeling can produce outcomes that magnify disparate behavior across groups [124].

Bias in Outcome evaluation

The outcome predicted by the learning task needs to evaluated against some measurements in order to establish the prediction quality. For a binary prediction, the overall accuracy can be a measure, whereas, for a continuous outcome, Mean Squared error (MSE) can be used. Other metrics such as sensitivity or specificity are also popularly employed. The choice of metrics against which the task is evaluated plays an important role in ensuring prediction quality and the model usefulness.

- Evaluation bias : When model performance fails to meet the goal of the task or models are compared against inappropriate metrics, they show evaluation bias [99]. For example, if the data sample contains 80 young individuals, 70% of whom were hired and 20 older individuals, 10% of whom were hired by the company, it shows that the data sample has skewed representation based on age. If the choice of metric is accuracy and the learning process picks up age as the discriminating factor, it is likely to predict young individuals as good for hiring. However, from the perspective of fairness, this measurement fails as there is discrimination in the outcome based on age. Moreover, this model may achieve an accuracy of 95% but hides the fact that the minority older population might have received the highest misclassification. Furthermore, when models are evaluated on inappropriate benchmarks or differ significantly from the target population, the evaluation is biased [124].

2.4.2 Detecting Bias and Discrimination

Fair ML brings together the disciplines of computer science, law and social science. The consequences are socio-economic, however, the heart of the fairness in machine learning lies at the intersection of the three fields. Discrimination discovery finds patterns of social discrimination in the data used for machine learning whereas discrimination prevention creates predictive models ensuring their decisions are fair [139].

For a machine learning model to be fair, it should be free from both direct and indirect discrimination. Direct discrimination does not exist when individuals or groups with similar sensitive features receive the same predictions. For example, consider two individuals Alice and

Bob, having similar skill set, education and working experience but differ only on on religious affiliation. Direct discrimination is removed when both receive same decision when they apply for a job, either accept or reject. On the other hand, indirect discrimination does not exist when individuals or groups receive different decisions that can be justified only on the basis of non-sensitive features. Consider again the same Alice and Bob, who differ on religious affiliation and also performance evaluation of work tasks. Indirect discrimination is removed when both receive different decisions, solely on the basis of performance evaluation. Mathematically,

Both direct and indirect discrimination must be discovered and prevented to make machine learning models fair. Discrimination can be measured on data, machine learning algorithms and the predictions made [139]. This section discusses four types of discrimination measures, namely statistical tests, absolute measures, conditional measures, and structural measures. The Table 3, lists the tests for each of the four types of discrimination measures. The first three measures can detect and quantify indirect discrimination, whereas structural measures detect direct discrimination.

Type of discrimination	Measures to detect discrimination	Example tests
Indirect	Statistical	Regression slope test Difference of means test Mann Whitney U test
	Absolute	Area under the Curve (AUC) Impact Ratio Balanced Residuals
	Conditional	Unexplained difference Belift ratio
Direct	Structural	Situation testing

Table 3: List of tests to detect and quantify direct and indirect discrimination in the data and predicted output.

Statistical tests only check for the presence or absence of discrimination in a dataset. As they do not quantify the discrimination, they are a quick way to demonstrate whether the dataset contains bias. Moreover, the tests can be applied to original or pre-processed data. Statistical tests have two main disadvantages. First, as they do not measure the magnitude or spread of discrimination, they cannot be used as standalone measures. Second, rejection of null hypothesis confirms presence of discrimination, however, if the null hypothesis cannot be rejected, it does not prove that there is no discrimination. Therefore, statistical tests cannot prove the absence of discrimination. Next, absolute tests go a step beyond statistical tests and measure the magnitude of discrimination between two protected groups. For multiple groups, each group is compared against the other. Like statistical tests, they can also discover discrimination only on the dataset. Therefore, they cannot be used as standalone tests to discover discrimination. Further, conditional measures go beyond the absolute measures by quantifying the differential treatment received by groups based on the features, excluding the sensitive feature. This implies that the remaining discrimination comes from using the sensitive feature. Lastly, structural measures quantify direct discrimination by identifying individuals affected by it.

2.4.3 Targeting Bias

There are multiple methods in literature to target bias at the three stages mentioned above. These can be identified as pre-processing, in-processing and post-processing corresponding to input data, predictive modeling algorithm and outcome evaluation. They are as follows:

Data Pre-processing
Techniques used to transform training data into a form that lower biases in the dataset come under the pre-processing stage. Many techniques have been proposed to unbias data. Sampling is a technique to improve representation of a sample of population in the dataset [1], [66]. The dataset can be stratified on the basis of each subset of the population, for instance on the basis of place of residence or age, and a defined proportion of random samples from each subset can be taken as their representative [1]. Preferential sampling is based on the idea that data samples that are close to the borderline tend to be wrongly classified and therefore should be preferred for higher sampling in the dataset [64]. This type of sampling can be justified in a hiring dataset because people from disadvantaged groups who are qualified for hiring form a small fraction in the dataset due to historical bias and are also likely to get misclassified or rejected as they are closer to the borderline compared to the people who are not qualified. Preferential sampling can improve their representation in the dataset, thereby, lowering historical discrimination.

Next, Simpson's paradox is a type of aggregation bias where observations in aggregated data disappear when examining unsegregated underlying data [119], which is important to handle in a dataset such as for hiring because of the complex interconnected patterns between the sensitive and non-sensitive attributes of people. A well-known example of Simpson's paradox is biased graduate admissions on the basis of sex [14]. It shows that while the aggregated data has bias against women, data per department does not reflect this bias because a large proportion of female applicants applied to departments with low admission rate for both sexes.

Another technique for discrimination discovery and prevention is by using causal models [120] to detect direct discrimination in the data by identifying the causal connection between the decision and sensitive attribute. Discrimination is discovered by partitioning the dataset into meaningful partitions based on sensitive attributes and discrimination is prevented by modifying the subsequent causal graphs or dataset [138]. Further, causal graph and causal modeling can also detect and remove discrimination at the individual level [96]. Causal models can help detect discrimination in hiring that is based on the sensitive attributes by identifying the attributes that directly of indirectly influence the hiring decision. However, causal modeling for hiring requires extensive domain knowledge.

Disparate impact is the unintended discrimination that produces different outcomes for different social groups. It can be partially removed while creating a trade-off between accuracy and fairness [45]. Disparate impact is unintentional, which is more aligned with the interests of law and policy makers than recruitment personnel of organizations. This is because the organization is not required to legally hire people who are representative of a fair society based on all sensitive attributes. However, policy makers will benefit more by studying the effect of providing or denying economic opportunities to different social groups.

Several other techniques exist in literature to remove bias and lower discrimination in the dataset, including massaging [25] and reweighing [65]. By massaging the data, some labels of the data are changed, which requires knowledge of the data distribution per label and sensitive attribute [66]. On the other hand reweighing keeps the data labels intact but changes the weights of data samples that are likely to be discriminated against, which is further used as frequency counts in the modeling phase [66]. Similar to massaging, reweighing also needs a careful study to understand which data samples need reweighing and by how much. Both techniques can prove useful in hiring despite the time intensive process. However, it is important to note that the meaning of sensitive attributes and primarily their distribution for new job applicants is bound to change. Massaging or reweighing techniques are limited in usefulness for dynamic datasets.

Predictive modeling Predictive modeling techniques modify algorithms by incorporating changes into the objective function or imposing fairness constraints on the loss functions to make the training discrimination aware.

Literature points to techniques that modify traditional algorithms such as Naive Bayes or Decision Trees. By training separate models, one for each value of the sensitive feature and combining them to create a classifier can produce a model that is not biased [26]. In the use of decision tree for fair classification, the changes are made to the splitting criteria [67], where the split that is least discriminatory with respect to the sensitive feature is used. These techniques are computationally demanding and no not adhere to any notions of fairness.

Predictive modeling can be altered for fairness by altering the initial conditions of the algorithm, thereby forcing a skewed training. An analogy is to reset the real world, so the algorithm trains to optimize what is not possible in the real setting. For instance, by first ranking the data samples in ascending order for sensitive group from negative class and in descending order for non-sensitive group in positive class and then flipping the class labels based on a measure of discrimination, an initial stage is set to train a classifier with a skewed and non-realistic data [65]. The goal is to produce a model which will make predictions based on this skewed learning, despite being trained on biased dataset.

Adding regularization, can be used to enforce fairness in the training phase. By adding penalty terms for over-fitting and violating some measure of fairness, the loss function can be penalized [70], for producing a model that would produce unfair outcomes. Such solutions may not be always be convex, thereby producing sub-optimal or local solutions.

Post - processing Post processing techniques adjust the trained model or further process the predictions to produce final outcomes that are not correlated with the sensitive attribute. The advantage of post-processing techniques is that it can easily operate on black box systems, which is the most common system available today [9]. When the inner working of the training phase is not understood or is too complex to modify, post-processing techniques can be the last option to remove discrimination from the pipeline.

Several techniques have been processed at this stage. One such technique requires only aggregated information about the data and uses only the sensitive attribute A, the prediction $\hat{Y}$ and the actual outcome Y, without any consideration of the relationship between X and A [59]. This post processing technique satisfies equal opportunity and equalized odds. Another technique relabels the predicted class of some leaf nodes made by the decision tree such that they do not belong to the majority class and also do not trade-off with accuracy [67]. Similar approach of relabeling has been used by models based on probabilities, by changing the class labels of the predictions close to the decision boundary [68]. Lastly, by changing the probabilities of the model, the probabilities of the decisions can be modified [26], which can be controlled to favor one decision over the other.

2.5 Towards Fairness Notions

The techniques proposed in section to target and address bias do not address the concepts of societal fairness. With the realizations that considerations of discrimination move beyond modification of data, a range of fairness notions have been proposed, which comes from ethical, moral and social understanding. For the purpose of fair machine learning, fairness as a social concept needs to be translated into a mathematical formulation. They help in identifying different social costs in a decision making system [9]. This section provides a mathematical overview of different flavors of fairness. The fairness definitions are based solely on the data or additional domain knowledge on data [101]. The definitions based on data come from studying

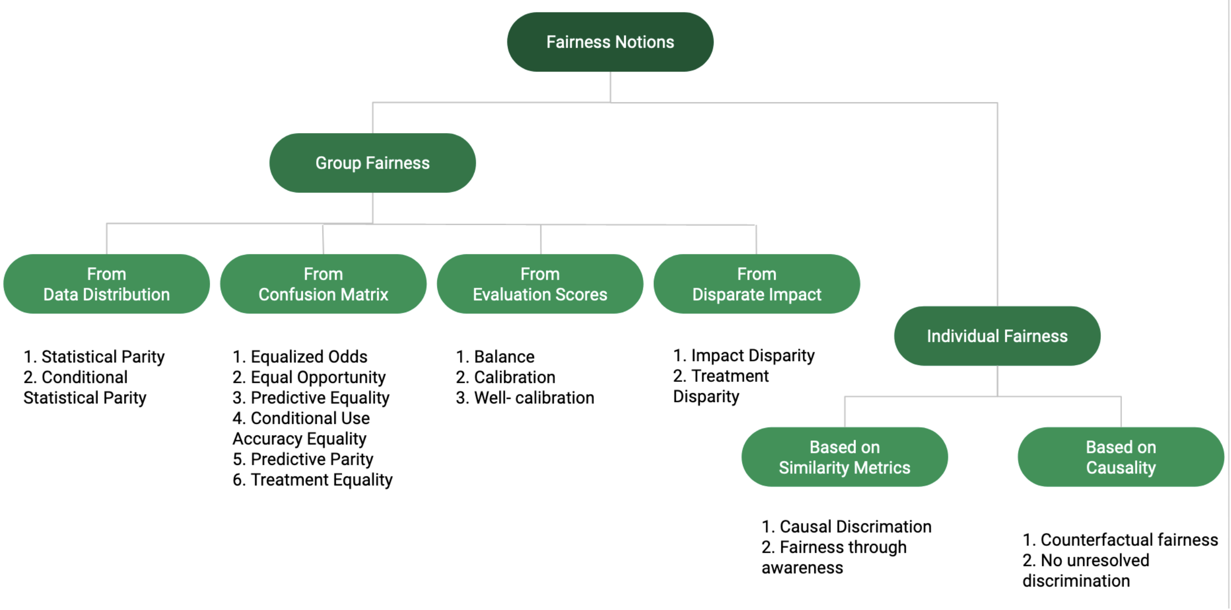


Figure 7: Classification of all the fairness notions

the data distribution, confusion matrix, disparate impact, evaluation scores, causality and similarity metrics as shown in figure 7.

2.5.1 Setup and Notation

This sections gives the notation and canonical formulation for mathematical formulations of fairness. The notations are consistent with literature. Consider a population indexed by $i = 1, ...n$, who are each represented by $v_i \in V$, , $y_i \in Y$ and $\hat{y}_i \in \hat{Y}$, where random variables V denotes all the features, Y denotes the actual outcome and $\hat{Y}$ denotes the prediction. The features can be subdivided into sensitive features A and non sensitive features X giving $V = (X, A)$.

For this discussion we keep A as a binary variable, where 1 indicates sensitive or protected or minority group and 0 indicates the non-sensitive or majority group. A social group reflects one group such as the minority group or the non-sensitive group. Y and $\hat{Y}$ are also binary random variables where 1 indicates the positive or selected class and 0 indicates negative or rejected class. $\hat{Y}$ is derived from the continuous real valued random variable S, which is a function of V.

Most fairness notions are defined using conditional probabilities. For instance $P(\hat{Y}|A = 0)$ represents the conditional probability of predicted outcome given the sensitive group. For example, in the context of hiring, we can predict whether a candidate i will be selected ($\hat{y}_i = 1$) or rejected ($\hat{y}_i = 0$) based on the features V_i containing non-sensitive X_i and sensitive A_i features. The features contain information about the education background, work experience, skill set, other performance measures and sensitive features such as age, sex, religious affiliation, etc.

2.5.2 From Data Distribution

The definitions of fairness based on the distribution of data depends only on the predicted outcome $\hat{Y}$ and not the actual outcome Y. There are two main motivations in favor of adopting these fairness definitions. First, when actual outcome Y can be uncertain or unknown (e.g the performance of someone not hired), it introduces uncertainty in error rate. Therefore, relying on the predicted outcome $\hat{Y}$, however biased it may be, can be considered more reliable [101]. Second, The relation between sensitive attribute A and observed outcome Y may represent the real world, but can be discriminatory. Therefore, it can be justified to rely on predicted outcome $\hat{Y}$, rather than the actual outcome Y. These definitions are also known as Independence condition $\hat{Y} \perp A$.

Statistical parity requires the prediction $\hat{Y}$ to be statistically independent of the sensitive attribute A : $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$. This means that equal prediction rates across groups regardless of the actual outcome Y [43]. Statistical parity is useful when one decision is preferred over the other. For a job applicant, getting selected is preferable compared to getting rejected. However, for a recruiter, both decisions are neutral as the goal is to only select a subset of applicants for the next round of interview process and finding the right and best candidate is important over the consequences of hiring or rejecting a candidate. This is also a suitable metric when there are legal requirements of equal acceptance rates for multiple sensitive groups. However, the downside of statistical parity is that it can be satisfied without satisfying fairness. For instance, to obtain equal acceptance rates for men and women in hiring, the recruiters can select qualified candidates from one group whereas, select only random candidates from the sensitive group to satisfy this fairness criteria. This results in masking [10], where random candidates instead of qualified candidates are hired in order to satisfy statistical parity.

Conditional statistical parity requires the prediction $\hat{Y}$ to be statistically independent of the sensitive feature A and additional feature say E: $P(\hat{Y} = 1|E = e; A = 0) = P(\hat{Y} = 1|E = e; A = 1)$. This means equal prediction rates across sensitive groups for same feature value $E = e$, regardless of the actual outcome Y [35]. This definition is a relaxation of statistical parity and treats similar individuals similarly. For example, among all job applicants with the same level of education, 50% of the Muslim men who applied get hired and 50% of Christian men who applied get hired. Similarity in this example is based on the same education level with religion being a sensitive feature. Conditional statistical parity is useful when another feature that is correlated with A needs to be prioritized for fairness (e.g. education level) and the difference in prediction across sensitive groups can be legitimately explained [98]. The downside of this notion is similar to the downsides of statistical parity - the condition can be satisfied by masking and without satisfying fairness.

2.5.3 From Confusion Matrix

The definitions of fairness based on the confusion matrix depend on both the actual outcome Y and predicted outcome $\hat{Y}$. Each cell in the confusion matrix asks for equality across sensitive groups. The values in a confusion matrix sum up to 1 across two rows and two columns, which implies that equality of a cell provides equality of its pair in the same row or column [101]. Some definitions are conditioned on the actual outcome, also known as Separation condition $\hat{Y} \perp A|Y$, whereas other definitions are conditioned on the predicted outcome, also known as Sufficiency condition $Y \perp A|\hat{Y}$.

Equalized Odds requires the positive prediction $\hat{Y}$ to be conditionally independent of the sensitive feature given the actual outcome Y : $P(\hat{Y} = 1|Y = y, A = 0) = P(\hat{Y} = 1|Y = y, A = 1)\ \forall y \in 0, 1$. This means that the prediction can depend on the sensitive feature indirectly through the actual outcome [59]. Therefore, people with the same actual outcomes are treated the same regardless of their sensitive features, which is from the perspective of the population being evaluated [101]. In the confusion matrix, this means that for $y = 1$, $\hat{Y}$ has equal True Positive Rate (TPR) across $A = 0$ or $A = 1$. Similarly, this means that for $y = 0$, $\hat{Y}$ has equal

False Positive Rate (FPR) across $A = 0$ or $A = 1$. Satisfying equalized odds , for example is when the probability of a person in hired class correctly predicted hired and probability of a person in rejected class being incorrectly predicted hired are the same for both men and women. This notion is suitable when the ground truth is available and the actual outcome Y can be trusted. Moreover, equalized odds is particularly useful when recall is more important than precision. This is a significant insight because an algorithm selecting candidates at the point of hiring needs to have high recall. Recall measures the proportion of candidates with actual outcome of getting hired are also predicted to be hired. This implies that True Positive (TP) should be high and False Negative (FN) should be low, thereby providing a high recall. Hiring is a recall sensitive scenario because cost of rejecting a qualified candidate is high as the organization will end up interviewing more unqualified candidates. Despite these advantages, equalized odds does not close the gap between sub-groups and is also difficult to achieve in practice [98]. Due to this reason, two relaxed variants of equalized odds exists, namely equal opportunity and predictive equality.

Equal Opportunity requires the positive prediction $\hat{Y}$ to be conditionally independent of the sensitive feature given that Y comes from positive class : $P(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1|Y = 1, A = 1)$. This means that the probability of being predicted in the positive class when the actual outcome is positive should not depend on the sensitive feature [59]. From the confusion matrix, this indicates equal TPR for both groups. Similarly, this also indicates equal FNR for both groups. An example satisfying equal opportunity is when equal proportions of people are selected from the qualified fraction of the sensitive group, such as men and women. This fairness notion is useful when the False Positive (FP) rate is not important. In practice this would mean that more unqualified employees also get selected for the next round in the hiring process, along with the qualified employees. This can be considered fair because it gives equal opportunity to all candidates, irrespective of the sensitive feature. This also implies that equal opportunity should not be applied when having high FP can have consequences (e.g firing a well-performing employee) [98].

Predictive Equality requires the positive prediction $\hat{Y}$ to be conditionally independent of the sensitive feature given that Y comes from negative class : $P(\hat{Y} = 1|Y = 0, A = 0) = P(\hat{Y} = 1|Y = 0, A = 1)$. It differs from equal opportunity because it checks for FP, whereas equal opportunity checks for TP. Therefore, predictive equality checks whether the accuracy of predictions is equal across sensitive and non-sensitive groups [35]. From the confusion matrix, this indicates equal FNR or TNR for both groups. Translating this to an example means that unqualified candidates are selected proportionately across groups. In the context of hiring, predictive equality is therefore, not a significant measure of fairness.

Conditional Use Accuracy Equality requires the actual outcome Y to be conditionally independent of the sensitive feature A given the same prediction $\hat{Y}$: $P(Y = y|\hat{Y} = y, A = 0) = P(Y = y|\hat{Y} = y, A = 1)$ $\forall y \in 0, 1$. This means the probability of actually belonging to the positive class for people who are predicted to be in positive class and the probability of actually belonging to the negative class for people who are predicted to be in negative class is equal [12]. From the confusion matrix, this notion requires equal Positive Predicate Value (PPV) and Negative Predicate Value (NPV) across groups. For example, if national origin is a sensitive feature with two values, native and non-native, where the latter is sensitive value, this fairness notion is satisfied when there is equal accuracy for native and non-native applicants from the predicted hired and rejected classes. Conditional Use Accuracy Equality is useful to be satisfied when precision is more important than recall. In the context of hiring, high precision implies that a low proportion of the candidates with the actual outcome of being rejected are wrongly predicted to be hired, which translates to low FP. Due to bias in the data indicating biased actual outcome, achieving low FP and thereby low precision is not important. This condition has a relaxation called predictive parity.

Predictive Parity requires the actual positive outcome $Y = 1$ to be conditionally independent of the sensitive feature A given positive prediction $\hat{Y} = 1$: $P(Y = 1|\hat{Y} = 1, A = 0) = P(Y = 1|\hat{Y} = 1, A = 1)$. This relaxation requires only equal PPV or precision across groups [32]. Similar to conditional use accuracy equality, this fairness notion is satisfied when there is equal accuracy for native and non-native applicants from the predicted hired class, where being a non-native is a sensitive value. This notion is insensitive to False Negatives (FN). Therefore, it is not recommended for hiring as low FN is desired.

Overall accuracy equality requires the prediction $\hat{Y}$to be same as the actual outcome Y irrespective of the sensitive feature A : $P(\hat{Y} = Y|A = 0) = P(\hat{Y} = Y|A = 1)$. This means same overall accuracy across groups [12]. For instance, if ethnicity is a sensitive feature, satisfying this fairness notion requires the probability of well-qualified applicants to be correctly accepted for the job and non-qualified applicants to be correctly rejected to be the same for both majority and minority ethnic groups. Overall accuracy equality is similar to conditional use accuracy equality but differs in the sense that it aggregates both positive and negative class misclassification with FP and FN [98]. This is not always desirable for fairness because cost of rejecting a qualified applicant from minority group can be justified to be higher than rejecting a qualified applicant from majority group. Therefore, this fairness notion can be impractical to satisfy at the point of hiring.

Treatment Equality requires the ratio of FN to FP to be equal for all social groups, irrespective of sensitive feature A : $\frac{FN}{FP}_{A=0} = \frac{FN}{FP}_{A=1}$. This simple fairness notion treats all groups equally by having the same proportion of misclassification [12]. For instance, if men have different ratio of FN to FP compared to women, then men and women are being treated differently by the algorithm. This fairness notion can be useful when only Type I or Type II misclassification is important for fairness. Treatment equality is limited in usefulness by itself but can be a helpful indicator to satisfy other fairness notions [12], and hence can be useful for hiring.

2.5.4 From evaluation scores

A typical regression problem in the ML pipeline predicts a real valued score instead of the prediction class. We present, three fairness notions based on this score. These notions are applicable to binary classification set-up as regression problems can be converted into binary classification problems.

Balance requires balance for the positive class or the negative class [76]. Balance for the positive class requires the equal expectation of score received by the positive class for all groups : $E[S|Y = 1, A = 0] = E[S|Y = 1, A = 1]$. This means that on average individuals from the positive class receive similar predictions irrespective of their social group status. For instance, all men and women with the actual outcome of being hired on average receive the same score during prediction. This is important when positive class is important and must have balanced prediction for the group. Similarly, balance for the negative class requires the equal expectation of score received by the negative class for all groups : $E[S|Y = 0, A = 0] = E[S|Y = 0, A = 1]$. This is useful when balance for the negative class is more important compared to the positive class. Both notions are not symmetric, which implies that achieving balance for positive class does not guarantee balance for the negative class. Balance for both the classes simultaneously has not been found in literature [98]. In hiring, fairness is required in who is selected and ultimately hired. Hence, balance of positive class is more desirable than balance of negative class.

Calibration requires positive outcome for the same predicted score irrespective of the sensitive feature : $P(Y = 1|S = s, A = 0) = P(Y = 1|S = s, A = 1)\ \forall s \in [0, 1]$. This means equal probability of actually belonging to the positive class for each predicted score $S = s$, irrespective of the sensitive feature [32]. For example, muslim women and christian women with predicted score of 0.8 should have equal probability of having their actual outcome of being selected, when religion is a sensitive feature. The same statement holds for every value of $s \in [0, 1]$. Calibration is similar to predictive parity but is a stronger notion of fairness because it remains valid for any chosen threshold to obtain a cut-off limit for candidate selection. This makes calibration suitable for scenarios where the threshold can be dynamic. A stronger bound on calibration is achieved by the next fairness notion called well-calibration.

Well-calibration requires positive outcome for the same predicted score S to be also equal to s, irrespective of the sensitive feature : $P(Y = 1|S = s, A = 0) = P(Y = 1|S = s, A = 1) = s\ \forall s \in [0, 1]$. This means that apart from satisfying calibration, probability for individuals to actual come from selected class is equal to the score for every given score [76] and this score reflects the probability of an outcome given all the features. For example, from the candidates who have a probability s of being hired, approximately s percent of them should have their actual outcome of being hired. This notion of fairness is suitable for hiring, however, it can be satisfied without any consideration of fairness. When there is enough representative data, A can be predicted from the remaining features and S can be approximated to $E(Y|A, X)$ which comes close to $E(Y|X)$, thereby satisfying well-calibration without any effort towards fairness [9])

2.5.5 Based on Similarity

As fairness notions are applied where human beings receive decisions, some fairness notions have been proposed with the consideration of the (dis)similarity between individuals based on their characteristics and sensitive features.

Causal discrimination requires the same prediction for individuals who have the same features but differ only with respect to their sensitive feature : $X_{a=0} = X_{a=1} \wedge A_0 \neq A_1 \Rightarrow \hat{y}_{a=0} = \hat{y}_{a=1}$. This means the same treatment is given to individuals who are similar but come from different social groups [49]. Causal discrimination looks at individuals and not how a minority group fares with respect to the majority group. For example, two individuals who have the same work experience, educational background and skill set but come from different religion are treated the same. This is a useful metric to understand the percentage of violations in the predictions for a dataset or get quantify how many people are treated unfairly by the model, which is useful for hiring.

Fairness through awareness requires the same prediction for any pair of individuals who have lower similarity distance than a given threshold [43]. For any two individuals i and j, where $P(X_i)$ is the probability distribution of over the outcome of prediction for i, D measures the distance between two probability distributions and d measure the similarity distance between the two individuals, fairness through awareness can be written as : $D(P(X_i), P(X_j)) \leq d(X_i, X_j)$. For instance, for a binary outcome, if the probability distribution for individuals i and j are [0.3,0.7] and [0.2,0.8], respectively, the distance between the distributions could be measured by, say Hellinger distance between them which is approximately 0.08. Here [0.3,0.7] means that the probability of belonging to the positive class is 0.3 and the probability of belonging to the negative class is 0.7. Now, if the similarity distance metric d between i and j is, say the euclidean distance between them, fairness through awareness is achieved if the Hellinger distance of 0.08 is lower than the euclidean distance between features of i and j. Though similar, fairness through awareness can be seen as a relaxed version of causal discrimination but provides a fine-grained analysis because it can also quantify how individuals are treated. A major challenge of this fairness notion is that a good collaboration is required between domain experts to define the similarity metric, which can be different based on the context and its requirements.

2.5.6 Based on Causality

Causal definitions of fairness go beyond mathematical constructs by incorporating additional constructs to help make value judgments [101]. These notions can help find pathways of unfairness and possibly bridge the gap between mathematical notions of fairness and the social concept of fairness.

Counterfactual fairness requires the probability for every individual with $A = a$ to get the same prediction, had the sensitive value been $A = a'$. It looks at the causal relationship between variables, rather than the statistical correlations between them [80]. For example, looking at the change in probability of receiving a positive or negative outcome for an individual with her ethnicity flipped, can show how the model is dependent on ethnicity for the prediction. Counterfactual fairness can also have multiple flavors, where instead of examining the effect of a direct dependence between sensitive outcome and predictive, an indirect effect can be examined. For example, sex of an individual may effect the education level they receive, which is in turn used to predict their hirability. This creates an indirect dependence between sex and hirability. The counterfactual scenario would be to set the education level, the person would have achieved if they belong to the opposite sex and examine the changes in prediction probabilities. This measure is useful to take the study of fairness beyond the boundaries of mathematical notions. It can also be used only with domain knowledge and professional expertise, as it is required to create causal graphs. Counterfactual fairness is difficult to achieve as it requires no change in prediction when the sensitive features are flipped as it is an unrealistic scenario.

No unresolved discrimination is achieved when there are no direct paths from from A to prediction $\hat{Y}$, except through a resolving variable [74]. From the example presented for counterfactual fairness, difference in prediction between sexes can be justified as it impacts education level, which can act as a resolving variable. This situation can be visualized as path A as not acceptable, and path B as acceptable. Whether such a justification is moral despite being under legal bounds debatable. This is a weak notion of fairness as discrimination can be masked by back-door paths [106], which can be intentional or unintentional if domain knowledge is missing or impact of the model is not understood by designers. No unresolved discrimination may not be beneficial in hiring, but other fields like medicine, where sex of an individual may be useful in predicting treatment.

2.5.7 Based on disparate treatment and impact

Legal concepts, namely disparate treatment and disparate impact, have led to the formulation of the mathematical fairness notions - treatment disparity and impact disparity [92]. Disparate treatment occurs when members of different social groups are given different treatment intentionally whereas, disparate impact occurs when the treatment differs across social groups unintentionally [10]. Algorithmic decision making can exhibit both forms of discrimination. It is important to note that algorithmic notions may not achieve the legal definitions of fairness, despite being inspired by them.

Impact Disparity requires same fraction of positive predictions across different groups [135], which in literature has been formulated as $p\%$ rule [15]. According to the $p\%$ rule, the proportion of positive predictions from the advantaged group, say q_a and the positive predictions from the disadvantaged group, say q_b should satisfy $\frac{q_b}{q_a} \geq \frac{p}{100}$. For example, if ethnicity is the sensitive feature and African-american is the sensitive group for an occupation and $p = 80$, then the number of African-americans selected should be greater than 80% of the non African-americans. This is a suitable fairness notion when the historical data and the outcomes present in it contain biases. However, if law does not help in choosing a fair p value, this notion can be satisfied without achieving social fairness. A related notion **Treatment disparity** requires the prediction, $\hat{Y}$ to be blind to the sensitive feature $A : \hat{Y} \perp A$. This means that the final prediction should not be based on the sensitive characteristics of an individual. While disparate treatment requires blindness to the sensitive feature, disparate impact takes it into account. A major disadvantage of disparate treatment as the sole fairness notion is that making an algorithmic decision making system blind to the sensitive feature can still produce unfair outcomes because of biases in the data, algorithm or evaluation process. It is a challenge to satisfy both these notions of fairness [134].

2.6 Granularity of Fairness Notions

Fairness can be defined at three levels of granularity. Instances of discrimination can thus be found by identifying individuals who are treated less favorably, discovering groups whose aggregate decisions are less favorable and sub-groups who are at a disadvantage despite their overall group receiving fair treatment.

2.6.1 Individual, group and sub-group fairness

Group Fairness also known as statistical fairness asks to treat different social groups equally [80]. After fixing a number of groups based on sensitive features, some chosen statistical measure needs to be valid across all the multiple groups [33]. Group fairness is easy to understand and obtain as it does not require any assumptions of the underlying population [73]. This level of analysis provides average guarantee for groups. One implementation checked the average weighted similarity of outcomes for two groups [11]. However, the average guarantees do not highlight the instances where some people are still unfavorably treated [33]. Moreover, some statistical fairness notions can be satisfied without reaching the social goals of fairness [73] and thereby, fail to provide meaningful guarantees [43]. Studies also show that multiple statistical notions are impossible to satisfy simultaneously [32, 76], making it difficult to operationalize group fairness.

Individual fairness asks for similar predictions for similar individuals [80]. A fine-grain analysis of fairness can be achieved by placing constraints on pairs of individuals [43]. A motivation for desiring this granularity of fairness is that people who are less qualified should not be preferred over more qualified ones [63]. Moreover, individual fairness provides better understanding of the social context than group fairness [33]. One implementation for satisfying fairness at individual level compares every individual of a protected group to every individual from another protected group and penalizes the algorithm for different decisions for similar individuals [11]. A distance or similarity metric is required to compare individuals to see how similar they are. The downside to this is that different metrics can produce different results. Therefore, knowledge of the data, especially, the relationship between features and labels are required to perform such fine-grain analysis [33], [73]. This also requires domain knowledge and understanding of organizational requirements.

Sub-group fairness is relatively new in literature and asks to satisfy group fairness notions over smaller collection of sub-groups [73], [72]. The motivation for this level of analysis is to pick best properties of both group and individual notions of fairness to get a fair outcomes over smaller-groups. The advantage of sub-group fairness is that group fairness can be satisfied while

	Independence	Separation	Sufficiency
Notation	$\hat{Y} \perp A$	$\hat{Y} \perp A \vert Y$	$Y \perp A \vert \hat{Y}$
Idea	Prediction should be independent of the sensitive feature.	Prediction should be independent of the sensitive feature given the actual outcome because the sensitive features may be correlated with the prediction. Correlation between the prediction and sensitive feature is allowed only to the extent it can be justified by the actual outcome.	Prediction is sufficient when sensitive feature and actual outcome are clear from the context because prediction is already based on sensitive feature, which is independent of the actual outcome.
Example	Skills required for a job should be considered independently of the sensitive feature such as religion or ethnicity.	The rate of incorrectly being predicted as rejected and the rate of incorrectly being predicted as selected should be the same for men and women.	The rate of correctly predicting positive decision over all the positive predictions and the rate of correctly predicting negative decision over all the negative predictions should be the same for applicants with or without disability
Cons	Can be satisfied with masking and no effort towards fairness	Unreliable if actual outcome is unreliable	Does not capture or limit additional dependencies between A, Y and $\hat{Y}$ introduced by the model [6]

Table 4: Independence, Separation and Sufficiency conditions

addressing intersectionality [33]. For instance, sub-group fairness can be satisfied for people belonging to two or more social groups. However, this level of fairness suffers from the same disadvantages as group fairness, albeit on a limited scale.

2.6.2 Independence, Separation and Sufficiency conditions

Most of the group fairness notions come from properties of joint distribution of sensitive features A, actual outcome Y and prediction $\hat{Y}$ or score S. Therefore, they are written as statements involving properties of these three random variables, which produces three groups of fairness criteria, namely Independence, Separation and Sufficiency conditions [9]. The notation, main idea, example and limitations of these conditions is listed Table 4.

It is seen from literature that independence and sufficiency are mutually exclusive and cannot be satisfied simultaneously [9]. The same hold true for independence and separation. Further, satisfying separation and sufficiency at the same time only results in degenerate solution [9]. This calls for fine grain analysis of different fairness flavors and granularity to understand which definitions are suitable in the context of hiring and discover meaningful trade-offs between multiple criteria for fairness.

2.7 Assumptions and Choices in Fair ML

We enlist several assumptions made and choices taken in literature when designing an algorithmic or prediction based decision making system. A wide variety of choices can be made regarding the statistical problem at the input, algorithm and output levels.

2.7.1 Dataset

The input data used to train the predictive model can produce different outcomes depending on the choices made. A dataset that will be used in the context of hiring contains measurements, categories and properties of employees or job applicants. The dataset may posses certain characteristics that makes it unsuitable such as poor quality of dataset, irrelevant or correlated features or bias [101]. A knowledge of the sources of the undesirable properties can help make decisions about how and when to handle them. Consider the scenario of predicting whether a job candidate will be a high performer. The dataset contains information about who was hired and was a high performer as well as who was hired but was not a high performer. However, the dataset has no information about the performance of candidates who were not hired. Such bias in the data should be taken into consideration when accounting for fairness [33].

2.7.2 Input features

The choice of features to be used, including sensitive features should be informed, motivated and consistent by the definition of fairness and the consequent affects that the model may have. There has been a debate on the inclusion of sensitive attributes in the input data, which most recently concludes that sensitive features be allowed in prediction systems. The work in [26] shows that removing the sensitive feature from the dataset does not remove potential discrimination due to red lining effects. This is because the sensitive features are correlated with other features and can be predicted, despite their removal. Another work [138] demonstrates damage to data utility when sensitive features are removed from the input.

2.7.3 Predicted outcome

The statistical problem assumes that by predicting a small number of outcomes, the broader goal of fairness defined at the policy level can be achieved [101]. For example, to predict sales target of a potential employee, to measure their suitability for being selected for a sale position. Careful choices need to be made about what outcomes are predicted because irrelevant or too few outcomes may affect decision quality, also refereed to as omitted payoff bias [27]. Next, the decision space for the statistical problem needs to be chosen. for instance, the decision space can be to accept or reject an applicant in the early stages of hiring. Furthermore, there are many choices to be made in the model building process such as type of model or its parameters, which has an impact on fairness.

2.7.4 Prediction algorithm

After handling the data and training a predictive model, it has to be evaluated. Three important assumptions are that first, outcomes for each individual can be evaluated as a whole such as accuracy or specificity, second, outcome of each individual is considered equal to the others, and third, each outcome can be simultaneously evaluated without temporal consideration [101]. Some of these assumptions might not lead to fair outcomes. The performance of a standard machine learning classifier can be measured by the errors in the prediction. If the performance goal in a fairness application remains the same - minimize the error in prediction - the model will fit the majority population [33]. For example, historically disadvantaged ethnic groups have lower access to education and resources and may achieve lower grades at university compared to the majority groups. When job performance distribution is this example is different for two groups, error rate may not be the best evaluation method.

2.8 Fair ML - Achieving Fairness Notions

The fairness notions proposed in literature have been applied using classical machine learning algorithms, causal graphs and neural networks. In this section, we present, the relevant technical applications of fairness notions.

2.8.1 Fair Classification

In a simple set up, the assumption is that a sensitive feature such as ethnicity defines a mapping to two groups - advantaged and disadvantaged. A classification task such as selection or rejection of candidates for a job interview is fair if the predictions are not correlated with the sensitive feature. The sensitive features that categorizes people into advantaged and disadvantaged groups is typically based on discrimination laws and can vary across regions and cultural contexts. People can have a combination of sensitive features and multiple group associations,say a 50 year old muslim black woman, which makes fair classification a complex problem to tackle mathematically. Different frameworks for fair classification are listed here.

To achieve fairness through awareness, with the idea of treating similar individuals similarly, fairness is proposed as an optimization problem where the goal is to find a relation from individuals to class outcome distribution which minimizes expected loss constrained by the Lipschitz condition [43]. The population is divided into two distinct groups based on their sensitive feature and the result obtained by solving the linear optimization problem shows that statistical parity is obtained, with a small loss in utility. This post-processing approach combines both individual and group fairness. However, similarity between individuals is based on a metric which can vary as per contexts. Other works achieve statistical parity by massaging the input dataset [65] or training a model for each sensitive feature [26].

Another approach to satisfy fairness notions is to reduce it to a set of finite linear constraints and learning an accurate classifier for the linear constraints [2]. This reduces the binary classification setting to a series of cost sensitive classification, where the the input contains costs for predicting each class along with the features. The costs are obtained from a reduced saddle point problem. The classifiers used for incorporating linear fairness constraints are logistic regression and gradient boosted decision trees. Fairness notions such as statistical parity and equalized odds have been satisfied with this approach and is an in-processing technique.

Equalized odds or equal opportunity can be obtained by learning a non-discriminating predictor in the post-processing step that minimizes the loss of incorrectly predicting the labels [59]. ROC curves are used to determine feasible space for a predictor and ternary search to solve the optimization problem. Results are shown for a Bayes Optimal Classifier on a banking credit dataset to predict loan defaulting. An improvement over this approach to produce a a nearly optimal statistical solution that is tractable to achieve equalized odds is shown in [130].

Calibration can be satisfied by obtaining calibrated probability estimates using decision tree and naive Bayesian classifiers [133]. The best modification to improve the probability estimates has been studied on a large unbalanced dataset that also contains misclassification costs. The decision tree is modified with smoothing by m-estimation and a variant of pruning, whereas the naive bayesian classifiers with binning. The methods are done in-processing or post-processing. Another approach to obtain calibrated output looks at how an existing classifier like Support Vector Machine (SVM) that does not output probabilities can be used for calibration [108]. This post-processing technique is done by separately training an additional sigmoid function and mapping the SVM's output to it.

To prevent disparate impact, convex margin based classifiers such as logistic regression and SVM can be used by incorporating to the $p\%$ rule in their objective function and constraints [135]. Another approach tries to remove disparate impact from the original dataset, thereby removing dependence on type of classifier chosen [45]. The motivation for this approach is that there is a relation between predictability of the sensitive feature value and disparate impact. Therefore, by using a designed 'rank preserving and repair' method, stratified dataset is created that can no longer be used to predict the original sensitive feature while maintaining class predictability. As the $p\%$ rule is difficult to incorporate in a classification system, it is translated into an equivalent problem - decision boundary fairness [135], confusion matrix [45] or difference in FP and FN rates [32].

A chosen fairness notion should ideally be independent of the classification algorithm as many algorithms should be able to implement a fairness notion [43]. Most of the literature looks at trade-off between classification accuracy and fairness notion specified and some even achieve relaxations of a fairness notion as it is difficult to fully remove discrimination. An overview of the literature shows that experimental results are obtained on only a single binary valued sensitive feature in a binary classification setting. In the context of job candidate selection, multiple sensitive features are present and can be categorical, real-valued or binary.

2.8.2 Causal approaches

Causal approaches are not completely data driven but need additional knowledge of the domain it is applied to as it clarifies how discrimination can propagate in the algorithmic decision making system. By representing the variables in a graphical form - features, sensitive features and target variables - it is possible to predict the effects of causes and infer counterfactuals. The motivation for using causal approaches for fairness is that the philosophies on justice incorporate same ideas; it is unfair to receive outcomes that is beyond an individual's control and therefore policies choose to eliminate treatment given to people that have causal dependence beyond their control such as nationality or sex [94]. In a causal graph, the nodes represent the random variables and a path is a set of connected nodes. If a node representing a sensitive feature, say sex is connected to a set of nodes that ultimately connect to the node representing job performance, then the algorithm trained on this data will be discriminatory. The causal graphs have been represented using structural equation model or bayesian network [102], [74], [94], [31].

No unresolved discrimination can be formulated as the presence or absence of a path from a sensitive feature directly or via a mediator variable to the predicted outcome. Using causal graphs to identify the discriminative pathways from the observed aggregated data distribution, this notion can be achieved by maximizing the likelihood constrained by the restriction of effects induced by paths [102]. Only a single sensitive feature can been considered in the experiments. It is also difficult to find instances of no unresolved discrimination when other fairness notions are considered [74].

Using causal pathways to remove discrimination in the dataset by correcting the data samples that produce different outcomes for similar individuals, who differ only on the sensitive feature [30], achieves counterfactual fairness. This correction is done during test time by changing the prediction that an individual would have received in a counterfactual world using the abduction-action-prediction technique [107]. The results are also demonstrated for a single sensitive feature.

Both counterfactual fairness and no unresolved discrimination provide a deeper understanding of how discrimination is propagated in a model as well as identifying instances of unfairness in an algorithmic decision making system. However, modelling these causal notions of fairness is computationally challenging.

2.8.3 Fair representation Learning

Representation learning tries to learn representations of data that contain all the informative features while removing noise or irrelevant features. Fair representation learning, then learns representations of input training data that contains maximum information to map to the actual outcomes but minimum information about the sensitive features. By obtaining a fair representation of the data, fairness can be handled at the dataset level before it is given as an input to any type of classifier. For example, by using a neural network based semi-supervised technique for clustering to find an intermediate representation of data that encodes all information but removes group membership details [136], fair representation can be learned at the dataset level. This approach maps every data sample to a probability representation in the same space by satisfying statistical parity, which is a group fairness notion. Further, by obtaining fairness at dataset level removes group membership information thereby, also satisfying individual fairness. The results are however, obtained for only a single sensitive feature.

Further advances in fair representation learning comes from two approaches, namely variational auto encoders and adversarial learning. The major advantage of using variational auto encoders is that the learned representation can prevent unfairness in further downstream tasks such as classification. By adding a regularizer to deep variational autoencoder architecture [75], which is already capable of separating existing features from the sensitive features, further separation can be achieved [95]. This approach is not based on any specific fairness notion and also considers only a single sensitive feature in the datasets. In order to achieve statistical parity by taking into account multiple sensitive features, fair representation can be learned [37], by disentangling [93] information from the sensitive features.

Adversarial learning where the predictor's ability to predict the target variable is maximized and the adversary's ability to predict the sensitive feature is minimized [137] can be used to obtain fair representation of input data. This techniques shows that a predictor and adversary can be learned to optimize any desired notion of fairness such as statistical parity, equalized odds and equal opportunity. Disparate treatment and disparate impact can be removed by learning a generator to produce fair data that is representative of the real data using generative adversarial networks(GAN) [51]. The generator and two discriminators play adversarial games to produce a generator that maximizes the utility, a discriminator that minimizes the ability to predict whether the data sample is real or fake and second discriminator that minimizes the ability to predict the sensitive feature [131]. Both techniques using adversarial learning to create fair representation show results for only a single sensitive feature.

The above mentioned literature primarily compares the techniques based on accuracy and a pre-defined discrimination metric. Discrimination metric is based on measuring bias with respect to the sensitive feature, which in some form measures the difference in proportion of positive classifications from sensitive and non-sensitive groups. Despite some of the approaches satisfying fairness notions, they lack clarity on how the fairness notions are achieved.

2.9 Why is Fair ML Challenging?

In this section, we identify five challenges in the field of fair ML.

2.9.1 Gaps between mathematical and social concept of fairness

Fairness is a social construct that cannot be measured mathematically- rather, it is measured based on observations or properties related to them such as difference in treatment, imbalance in distribution of harms or benefits or socioeconomic consequences. This creates a mismatch between what is required to be measured and it operationalization [62]. Each fairness notion is based on satisfying a social understanding of fairness but they can often fail to satisfy their social counterparts. For example, the legal concepts of disparate treatment and disparate impact have given rise to two mathematical notions of fairness, namely treatment parity and impact parity, respectively. However, it may be possible to satisfy the mathematical notions without satisfying its legal equivalent [92], which creates a gap between legal and algorithmic definitions of fairness.

A study [121] was conducted to understand people's perception of fairness and map it to their mathematical counterparts by asking participants to judge if a fair algorithm's predictions were fair. It found that most people consider statistical parity as fair and have difficulty understanding complex definitions and they tend to choose accuracy over fairness when fairness is in a high stakes context such as predicting life expectancy. The study also concludes that most people prefer more context on how decisions are made so that they can make better and informed judgment of fair algorithm's outcomes.

2.9.2 Accuracy-fairness trade-off

Most works in fair classification, satisfy a fairness notion as an additional constraint in the optimization process, which brings a fairness while reducing accuracy. A study [11] demonstrates using logistic regression for group and individual fairness notions that accuracy reduces when fairness increases. It also shows that the choice of fairness notion and the dataset or domain can result in different accuracy fairness trade-offs. Optimization for accuracy and equalized odds has been shown to be an NP-hard problem [130]. In the domain of public safety, for instance, it can be challenging to satisfy existing fairness notions [35] while achieving high accuracy. Satisfying fairness notions might set free more high risk defendants, which is caused by reduction in accuracy.

To reduce the trade-off between accuracy and statistical parity, a study [46] looks at three machine learning algorithms namely, logistic regression, support vector machines and adaptive boosting. They show that by shifting decision boundary for the least error for statistical parity, the gap is the lowest when compared with random relabeling and massaging. Another work [72] demonstrates that the trade-off between accuracy and fairness is low when sub-group fairness is considered. Lastly, another work [53] shows that it is possible to either achieve fair outcomes or high accuracy but not both, through a study on COMPAS data.

2.9.3 Many fairness notions cannot be satisfied simultaneously

An algorithmic decision making tool called COMPAS has received widespread attention and criticism in recent years because it used features such as such family or neighborhood correlated with race to predict the risk of criminal recidivism in the U.S [4]. It satisfies two sufficiency conditions - Positive Predicate Value (PPV) [40] and calibration within groups [47]. This means that the proportion of defendants who got rearrested from the high risk group is same regardless of race and both White and Black defendants had equal probability of being high risk for the same predicted score. However, the tool did not satisfy equal False Positive Rate (FPR) by race [4], which is a separation condition. Therefore, despite satisfying two notions of fairness, the decisions were unfair because Black defendants were predicted with higher risk of recidivism than they actually were also had higher rate of misclassification as high risk compared to their White counterparts [4].

To satisfy both FPR and PPV, which come separation and sufficiency conditions, respectively, requires either the distribution of predicted scores to be same across groups or that some groups never a certain outcome [101]. In the context of hiring, say for male and female candidates, both groups have same base rates for selection or one of the groups is never rejected, which is an unrealistic scenario. Along the same lines, the notions of balance for positive class, balance for negative class and calibration cannot be satisfied simultaneously, because it requires equal base rates for all the groups or a prefect prediction of score s as 0 or 1 only [76], [32], which is also not realistic. These three fairness notions - positive class, balance for negative class and calibration - was also not satisfied for COMPAS tool. Moreover, calibration cannot be satisfied along with equalized odds [76], but a relaxation of equalized odds can be satisfied with calibration [109] at the cost of increasing misclassification for one social group. Lastly, statistical parity cannot be satisfied along with any fairness notion from separation or sufficiency condition [9]. These insights point towards the need for meaningful trade-offs between different fairness notions and their implications.

2.9.4 Difficult to reconcile group and individual fairness notions

When comparing group and individuals notions, literature shows that the former has been given more attention and has been preferred over individual fairness [82]. The three conditions of independence, sufficiency and separation are all group fairness notions, which contain an extensive list of notions, whereas there are comparatively fewer fairness notions targeting fairness at individual level. The point of consideration is that it is possible to satisfy a group fairness notion while still treating some individuals unfairly [17].

Some studies have looked into satisfying both levels of fairness. One approach to satisfy both group and individual fairness looks at creating a fair representation of the data from the existing data by removing group membership information while encoding all other relevant information [136], thereby removing information that can distinguish individuals based on sensitive characteristics and also attain fair outcome for the entire population. Another approach has created a new notion called sub-group fairness [73], with the idea of satisfying group fairness on sub-groups rather than on individuals. This approach satisfies statistical parity and equal opportunity for the sub-groups. Lastly, political and philosophical arguments have been made about why two mathematical fairness notions need not be conflicting [17].

2.9.5 No clear guidelines on how to operationalize fairness for given domain

There are tensions between different fairness notions and at what granularity they can be achieved. Without an understanding of the working or formulations of different fairness notions, it can be difficult to incorporate them for any context such as hiring, admission or lending. Most people when faced with multiple notions of fairness, tend to choose the simplest - statistical parity [121] and need more context on how decisions are made to better understand fairness. In the domain of financial lending, a study has experimented with different fairness notions and recommends separation condition as appropriate for credit scoring [79] because it is sensitive to misclassification cost, which is important for lending. While another study [115] found that ordinary people prefer calibration fairness to be most fair in the context of loan decisions.

There is no consensus on which fairness notions to operationalize for a task or how to measure the fairness of the algorithmic decision making system. This makes it difficult to adopt fair practices and policies for fair ML [139]. Some choices that can help set up guidelines are the availability, presence of features correlated with sensitive features, importance of precision versus recall, importance of FP versus FN, misclassification costs, prediction thresholds, intersection between sensitive features, legislation and organizational policies [139].

2.10 Additional challenges of Fair algorithmic selection of job candidates

While the challenges of fair ML identified in the previous section remain, applying fair ml to the field of hiring brings additional challenges. We first, present on overview of algorithmic hiring and then discuss its various limitations and challenges.

2.10.1 Overview of Algorithmic hiring

Algorithms are becoming a helpful source of advice in recruitment, selection and other HR activities. Recruitment and selection are distinct because algorithmic technologies are used for recruitment, which finds candidates and selection, which selects candidates from the found candidates [78]. Recruitment is done using search engines and recommender systems which can be based on say, collaborative filtering to find well-fitting candidates or online advertising [29],[24]. Here, algorithmic decision making helps recommend candidates to recruiters or jobs to candidates. Selection on the other hand, is done using algorithmic analysis of CV, video resume or job application, which can be based on say, text mining or Natural Language processing (NLP) [132], [81] or Facial Expression Processing (FEP) [103], [55]. Algorithmic based decision making system for both recruitment and selection, also known as sourcing and assessment is widely used in industry [89].

While research continues in the field of Fair ML, there is a growing interest on its applicability in the context of hiring. One approach looks at group fairness by finding partitions of candidates in a space defined by a combination of sensitive features [44]. It uses a decision tree to make the split on the sensitive feature that results in the most unfairness, thereby showing which candidate group if selected, leads to highest unfairness. The results are obtained on a simulated dataset from crowd sourcing platform. Another study [122] confirms the presence of gender bias in online hiring on a hiring platform, TaskRabbit where individuals or groups can find a list of candidates for small duration tasks such as gardening or painting. When fair ranking algorithms are applied, it improved the representation of underrepresented candidates only when they had similar features to the majority. It also looked at adverse impact at the top 4 ranks using counterfactual fairness and found that their fair ranking algorithm prefers men at highest ranks compared to women. Next, inspired by domain experts who manually select a linear combination of weights to predict job performance while ensuring fairness, [50] uses Evolutionary multi-objective optimization (EMOO) to find algorithms that jointly increase job performance and reduce adverse impact. The goal of the obtained algorithm is fairness and legal compliance when transforming a list job candidates into a ranked list based on a linear combination of weights. The results are obtained on three data sets obtained from SIOP 2018.

Most of the commercially available tools providing algorithmic decision making for recruitment and selection do not disclose how decisions are made. A study [111], looked at 18 vendors who provide algorithmic decision making at these stages to understand how they work based on publicly available information. It was found that only 15 vendors vaguely mention bias and only half of them discuss legal compliance. Moreover, most vendors do not explain which biases are removed, how are they removed or how models are validated.

2.10.2 Limitations and Challenges

Along with the challenges of Fair ML discussed previously, fair hiring introduces additional challenges. They are:

- **Intersectionality** When looking at sensitive features - a vast majority of literature reviewed has focused on a single binary feature. However, people have multiple sensitive features, which can overlap and can be multi-valued. For example, the three sensitive features - race, sex and age are all different types such as categorical, binary and numerical. Job applicants have an overlap of sensitive features which brings an additional layer of complexity while incorporating fairness. A study [23] on three commercial vendors for gender classification based on facial images showed that dark skinned females were the most misclassified and light skinned male subjects had high accuracy in classification. One of the first literature [100] advocating adoption of intersectionality in selection problem shows that good utility can be achieved with intersectional constraints.

- **Interpretability** If recruiters who algorithmic decision making systems, need to explain why they can continue using the system, they need to understand the working of the system. There is a need for interpretability as employers and recruiters should have the possibility to understand the inner workings of the algorithmic process [88]. An important aspect of this is the need for accountability for decisions and transparency to applicants and employers in order to make fair decisions [38]. A study [111] to understand interpretability shows that knowledge of 5 factors namely, features, target variable, training data, validation and fairness notions, is needed to make the process more transparent and interpretable.

- **Biased dataset** For hiring, the predictions indicate employee performance, suitability, organization fit or a similar metric. The data contains this information only on who was hired; there is no performance data for someone who did not get hired. Predicting candidate job performance comes from biased data. Furthermore, many qualified applicants do not apply for jobs because of various reasons such as conflicting values, identity mismatch or personal concerns. This removes their representation in the data.

- **Limited correctional measures** It is only possible to observe false positives - those who get selected but do not get hired or perform well when hired. There is no data on false negatives - those who did not get selected but would have performed well. This means that if some social groups always get excluded from the selection process, it creates a high representation of candidates from certain social groups who never get hired. It creates difficulty in identifying this issue and lowers the possibility of correcting the dataset.

- **High cost of data procurement** Any ML task requires high quality dataset, which holds true for fair hiring. Data collection for this domain is expensive because- 1) it needs to be collected from large organization with a huge employee base, 2) the organization needs to monitor and store data related to its job applicants and employee performance, 3) the organization is not legally allowed to share information about its employee and 4) privacy and ethical concerns with the risk of identifying individuals and causing harm.

2.11 Human centered approaches in Fair ML

Recent literature shows the focus shifting towards human centered approaches to fairness conducted through online surveys [121], [115], [18], [84], [41], [54], [71], face to face interviews [129], [87], [89], [86], or participatory design approaches such as workshops [129], [22]. While some studies have looked at understanding how lay people perceive mathematical fairness through different dimensions of such as justice, emotions evoked, trust and attitude, others have similarly looked at the perception of fairness by people personally affected by algorithmic decision making. In this section, we discuss studies on lay people's and affected communities' perception of fairness, how explanation styles affect their judgment and recommendations for practitioners. We conclude this section, recommendations for practitioners.

2.11.1 Lay people's Perception of algorithmic fairness notions

To understand lay people's perception of fairness, [121] surveyed people to find out which fairness notions is more discriminatory in three contexts of recidivism prediction, skin cancer diagnosis and different level of impact on individuals, where the data contains 2 sensitive features - race and gender. They found that people have difficulty judging complicated fairness notions and tend to choose statistical parity. Moreover, people prefer high accuracy than equality for a high stakes scenario but asked for more context on how decisions were made. While [121] studied the fairness notions namely, statistical parity, error parity, equality of false positive rate and false discovery rate, [115] studied three other fairness notions namely fairness through awareness, calibrated fairness and never favor a worse individual over better one. By presenting scenarios in the domain of lending, the research investigated lay people's perception of fairness and how it changes when an additional sensitive feature is presented. They conclude that calibrated fairness was the most preferred notion in the context of lending.

Another research [54] studied how people perceive the usage of different input features, including sensitive features for making recidivism prediction decisions. They found that there is no consensus in the fairness of input features used, unfairness concerns of people go beyond discrimination and features can be considered unfair to be used despite not being a sensitive feature such as education, quality of life or involvement with family. Lastly, to understand the fairness judgment and the emotions evoked, [84] surveyed people in the context of managerial task to find that people consider algorithmic decisions as fair for mechanical tasks such as assigning work or scheduling, whereas they think algorithmic decisions are less fair than humans in tasks such as selecting candidates for a job or evaluating performances. They conclude that people's knowledge of the decision makers has an influence on the validity and trust in the decision made.

2.11.2 Affected people's perception of algorithmic fairness notions

Moving from ordinary people's perception of fairness, [129] and [22] investigate how people affected by algorithmic decision feel through face to face interaction with people. The research from [129] interviewed people from marginalized communities and found that algorithmic outcomes evoked negative emotions regarding racial and economic injustice and reduced their trust of the products of companies using algorithmic decision making. Furthermore, the research from [22] interviewed multiple stakeholders in child welfare service and found that people have many concerns with the system, the different scenarios in child welfare service and the degree of reliance on algorithmic decisions. The research also highlights affected people's concerns with bias in data and biased decisions made by the human relying on algorithmic decisions. Lastly, [87] uses participatory design by interviewing multiple stakeholders to obtain best practices to design an algorithmic decision making service to allocate donation to non-profit organizations. They conclude that different stakeholders have different fairness motivations, which depends on people, context and the type of interaction they will have with such systems.

2.11.3 Effect of explanation styles for algorithmic decisions

By surveying people on 5 common situations through fictional scenarios such as loan, work promotion, car insurance, flight re-routing and freezing of bank account, [18] conclude that there is no best way to explain algorithmic decisions to people. The reason being that explanation styles does not effect how people respond to (in)justice given to the fictional scenarios in the 5 situations. People's concerns with algorithmic decision making include lack of human involvement, reasonable outcome and morality. However, the study [41] shows that people consider some explanation styles unfair while some explanation styles increase their trust of the system in the context of recidivism prediction, where the severity of the decision is higher than the common situations in [18]. Moreover, people's prior position on algorithmic fairness has an influence on their reaction to algorithmic decisions.

Another related work, [117] studies how trustworthiness of an algorithmic decision making system changes when ordinary people are given varying levels of information about the decision making process and it's outcomes. Based on an online survey in the context of home loan approvals, they found that more information of the underlying system increased people's perception of trust in the system, however, it slightly decreased when shown the importance of input features such as education, gender and marital status. This can be connected to the results in [54], where people's perception of fairness in the outcomes decrease when shown the usage of input features such as education and family background, apart from race.

2.11.4 Algorithmic hiring in Practice

A study [111] examined what the vendors of algorithmic pre-assessment of job candidates disclose about their products in terms of features used, target variable, training data, validation and fairness notion. Based on publicly available information they found that assessment is available in 3 main forms - questionnaire such as personality test or situational judgment test, video interview analysis and gameplay. The target variable and training data is either selected by the vendor or the client's past and current employees, which can also be customized for job roles or client requirements. Most vendors, however did not disclose their validation methodology. Moreover, most vendors only check for legal compliance such as disparate impact or removal of variables correlated to sensitive features to reduce discrimination.

Another study [89] interviewed HR professional's who use algorithmic decision making systems for finding and assessing job candidates and reveals that HR professional's rely on the outcomes and the scores given to the candidates by the algorithmic decision making system to filter out a large number of applicants and shift their resources to make selections from the system's outcome. While some HR professionals onboard the tool after legal consultation, only few others consult I/O psychologist, approve the features, validation methodology or fairness notion used.

The implication of this is two fold. First, it limits the HR professionals' ability to improve diversity within their organizations because of low control over the training data or fairness notion used. Second, it limits the organizations using such tools from fully understanding the implications of the tool, justify its use and meet their goals of diversity and inclusion.

2.12 Recommendations for practitioners of Fair ML from Qualitative Research

- **Design fairness as a value in product design** The research [86] ran several experiments with groups of people who have some background or knowledge in fields adjacent to computer science and found that a significant proportion (33%) participants think that algorithms are less than fair, when the outcomes affects their groups or them individually in the context of fair division algorithms. They recommend accounting for social and human behavior in designs which can be difficult to express mathematically.Further, [129] recommends designing fairness as a value in product design, which could include diverse training data, diverse target group, ethical considerations and awareness of implications of various choices in designing the system.

- **Transparency of the process of the decision making system** While surveying students in fields related to algorithmic development, [71] found that students who agreed with the outcomes did not necessarily think that the outcomes were deserving and appropriate input features did not necessarily correspond to being fair. They recommend transparency in the input features used, including sensitive features and the procedures followed while making decisions. Another research survey [117], concluded that informational fairness can be achieved by informing users about which input features and how they are used in the decision making.

- **Participatory design to account for diverse perspectives** The study [129] conducted informative workshops and interviewed people from marginalized groups to find that many members from marginalized communities expressed anger or disappointment when decision making systems contain social prejudice against their communities. They recommend accounting for diverse perspectives, including those of marginalized group and engaging with community groups to collaborate on designs. Such approaches can also ensure that items relevant to under represented communities stay in context when designing systems [61].

  Participatory approaches were used by studies in [22] and [87], where various stakeholders' in a child welfare system and a food allocation organization, respectively, were interviewed. This approach was used to elicit different concerns, values and expectations of affected stakeholders'.

  While [22] and [87] examined the importance of involving various stakeholders in the design process, the research [118], on the other hand studied one approach of allowing community members to deliberate different machine learning models to pick the one most suitable for their goals and values. This type of participatory design approach allowed members to weigh different trade-offs and facilitate informed decision making with a final group proposal.

2.13 Research Gap

From this survey, we see that various fairness notions have been proposed in Fair ML literature, which has been applied in many domains such as financial lending and recidivism risk prediction as data-driven automated decision-making systems. Despite their usage, Fair ML has many open challenges that need to be addressed to ultimately handle the biased and discriminatory decisions that impact people.

Applying the fairness notions using these automated decision-making systems brings additional challenges in hiring. Commercial solutions to aid hiring help source or pre-assess job applicants. However, the decisions reflect many biases. Moreover, there is evidence of vendors and companies offering and using algorithmic decision-making for recruitment and assessing candidates, which lacks transparency on fairness notions, feature set, target variable, training, and validation data. The reliance of HR professionals on the decisions by such systems prevents them from understanding and justifying the use of these decision-making systems. There is no research on how HR professionals and organizations who use Fair ML, understand and agree with the fairness notions used behind the algorithms they use.

Recent literature points to several reasons why qualitative research on human perception of fairness is needed. First, ordinary people's judgment and perception of fairness show that one fairness notion cannot be used for different contexts. Second, people affected by automated decisions, when explained the process and concepts of automated decisions, are concerned about potential bias and discrimination, which lowers their trust in such systems. Lastly, results from participatory research show that the concept of fairness differs for different stakeholders.

This research aims to gain insights into the alignment of organizational representatives with different mathematical notions of fairness in terms of their understanding, impression, and perception of fairness towards achieving their goals of employee diversity through early candidate selection in hiring.

3 INITIAL DOMAIN INTERVIEWS

To aid the study design and to test the practicality, feasibility and research directions of our research, we we conducted initial domain interviews. We conducted semi-structured interviews with five participants. In this Chapter, we first describe the participants and provide background descriptives in section 3.1. We then describe the topics of the semi-structured interviews in section 3.2. The findings from the semi-structured interviews are described for each topic separately: current candidate selection process, the utility for different stakeholders during early stages of candidate selection, how fairness, diversity and inclusion are handled in the current selection process and how fair selections can be made. We provide these findings and a table with aggregate responses in section 3.3.

3.1 Participants

We conducted semi-structured interviews to understand the candidate selection process, current ambitions of HRM professionals regarding questions of fairness, and diversity current and future candidate selection process. They were recruited by direct emails obtained from connections of the research group. A total of 5 professionals participated in the semi-structured interviews, out of whom 4 are female and 1 male, with working experience in the field ranging from 0.5 to 23 years. All the 5 participants work in the Netherlands and none of the participants or their organizations have used automated decision making in the context of early candidate selection in hiring. The participant information is shown in Table 5.

Participant ID	Job Title	Experience (in years)	Gender
PS1	Professor, Psychology in hiring practices	20	Female
PS2	HR Advisor	5	Female
PS3	Senior Talent Acquisition	23	Male
PS4	Talent Acquisition specialist	4	Female
PS5	Admission Coordinator, Bachelor Programme	0.5	Female

Table 5: Background information of the Participants of the Pilot Study.

3.2 Semi-structured Interviews

Each participant was interviewed independently. The participants' informed consent was obtained by presenting them with an opening statement containing the goal of the interview, information to be collected and the contact details of the researchers. The semi-structured interviews covered 4 broad topics as follows:

- How does the current candidate selection process work?
- What goals do the different stakeholders strive for during recruitment and early stages of candidate selection?
- How is fairness, diversity and inclusion handled in the current selection process?
- How can fair selections be made in the early stages?

Each broad topic consisted of sub-questions, to which participants were allowed to respond openly. Written notes were taken during the interviews as the interviews were aimed to inform the study design.

3.3 Findings

The aggregated responses for the 4 broad topics obtained by analyzing the interview notes through discussions within the research group are shown in Table 6.

3.3.1 How does the current candidate selection process work?

The early selection process is performed by the vacancy holders who create the job vacancies and selection committees appointed by HR functions. The main role of the HR functions is to ensure transparency of the process. The selection makers have the option to explain their choices and decisions about each candidate they review. However, they are not required to provide an explanation. Most selection makers often make notes about candidates. Most organizations, according to the participants do not have well-formulated explicit diversity criteria during selections.

3.3.2 What goals do the different stakeholders strive for during recruitment and early stages of candidate selection?

The goal for different stakeholders such as the organization, HR functions and selection makers differs. The organization looks for candidates who make the organization's population diverse and enable the company's growth, whereas HR functions help find a diverse applicant pool for the different job vacancies in the organization. The goal for selection makers comes down to finding candidates with the right skill set, motivation and capability to do the job well.

3.3.3 How is fairness, diversity and inclusion handled in the current selection process?

Organizations take various measures to operationalize fairness and diversity in the selection process. They begin by creating textual descriptions of job vacancies that is likely to attract a diverse applicant pool. Internally, organizations may also provide bias and non-discrimination training to the selection makers and HR functions. Legal compliance is also taken into consideration. Many organizations also have diversity in their statements of mission, which may or may not be known by selection makers. Organizations are allowed to access some sensitive features of the applicants. Some use gender as a factor while making early selections but do not have policies or directions on monitoring other sensitive features.

3.3.4 How can fair selections be made in the early stages?

All participants identified gender as an important sensitive feature, while some also indicated nationality and ethnicity as important. However, none were able to coherently argue about the need to monitor different sensitive features. The confusion was caused by legal limitations and company policies on the use and access of the sensitive features. Further, every participant stated that it was impossible to monitor False Negatives in the selection process as per GDPR (General Data Protection Regulation) guidelines because applicant information cannot be retained by the organization with the applicant's consent. However, all participants indicated the possibility of monitoring for False Positives in the selection in the future.

Topic	Sub-questions	Aggregated responses
Current selection process	Who makes the early selections?	*Usually people who create the job vacancies or selection committees. It can also be based on test scores obtained.*
	Role of HR in selection process	*To oversee the process, appoint selection committees and screen candidates for hard requirements.*
	Are selection decisions explained?	*It is optional. Selection committees can be asked to explain their choices.*
	Are there diversity related criteria during selection?	*Rarely seen in practice. Some organizations work on it more than others.*
Goal of stakeholders	Organization	*Organizational goals such as diversity, motivation, reputation, company growth, job performance, organizational fit, etc*
	HR	*Finding a diverse applicant pool for vacancies.*
	Selection makers	*Skills, knowledge, experience, capability, motivation, etc*
Diversity and inclusion in current selection process	Operationalization	*1) Adhere to legal legislation to prevent discrimination. 2) Edit Job Vacancy description to attract diverse applicants. 3) Train committees and selection makers on diversity.*
	Awareness of mission statemenent on diversity	*Most employers making the selection are aware of the statement but are not always capable of putting it into practice.*
	Acesss to sensitive features	*Yes, only within the organization.*
	Sensitive features or groups monitored	*Gender and sometimes nationality.*
How to make fair selections	Important sensitive features	*Gender, nationality, ethnicity.*
	Should senstive features be monitored?	*Confusion over its legality.*
	How to monitor False Negatives?	*Not possible. As per GDPR, applicants information cannot be retained without their consent*
	How to monitor False Positives?	*Can be monitored by identifying the applicants who get rejected in the later stages of selection.*

Table 6: Findings of the Pilot Study.

These findings helped formalize the study design. First, the domain interviews helped narrow the scope of the research to organizational representatives who have involvement and influence over the early selections in hiring. Second, gender is the only sensitive feature that is feasible and legal to track by organizations, which will be the sensitive feature of focus in the study design. Third, the interviews showed the feasibility and scope of conversations with further domain experts regarding concepts of fairness.

4 METHODOLOGY

In this section, we describe the formulation of the study design and the methods adopted to answer the research questions. The study design (4.1) describes the relation to a binary classification set-up, the six chosen mathematical fairness notions and design of the user-friendly explanations and illustrations. The methods (**??**) describe the recruitment of the research participants (experts), how and when the they were interviewed, the interview process and the analysis performed. This methodology was approved by the Human Research Ethic Committee of the Delft University of Technology.

4.1 Study Design

4.1.1 Mathematical Fairness Notions

Formally, all of the notions can be implemented as optimization constraints in a binary supervised machine learning classification problem. In this problem, we have a collection of N candidates, where each candidate is represented by a given feature set X, a single binary sensitive feature, A and true outcome, Y, which is a binary variable indicating whether a candidate was selected to advance to the next stage or not. A classifier makes predictions $\hat{Y}$, that should be as close as possible to Y, while the fairness notion of interest is being satisfied.

Fairness Notion	Motivation	Definition	Useful when	Disadvantages
Fairness through awareness (FA)	Similar people should be treated similarly	A distance metric such as hellinger distance between the class probability distribution of two people should be lower than the distance between their features	It is important to quantify how individuals are treated	The similarity metric needs justification based on context and requirements
Counterfa fairness (CF)	Study of fairness can go beyond the boundaries of mathematical notions and should be done with domain experts. It is important to look at causal relationship between variables, rather than the statistical correlations between them	Probability for every individual with A=a to get the same prediction, had the sensitive value been A=a'	Domain and expert knowledge is available and it is important to show instances of unfairness that can be understood by humans	Difficult to achieve in practice

Table 7: The 2 individual fairness notions - Fairness through awareness [43] and Counterfactual fairness [80]

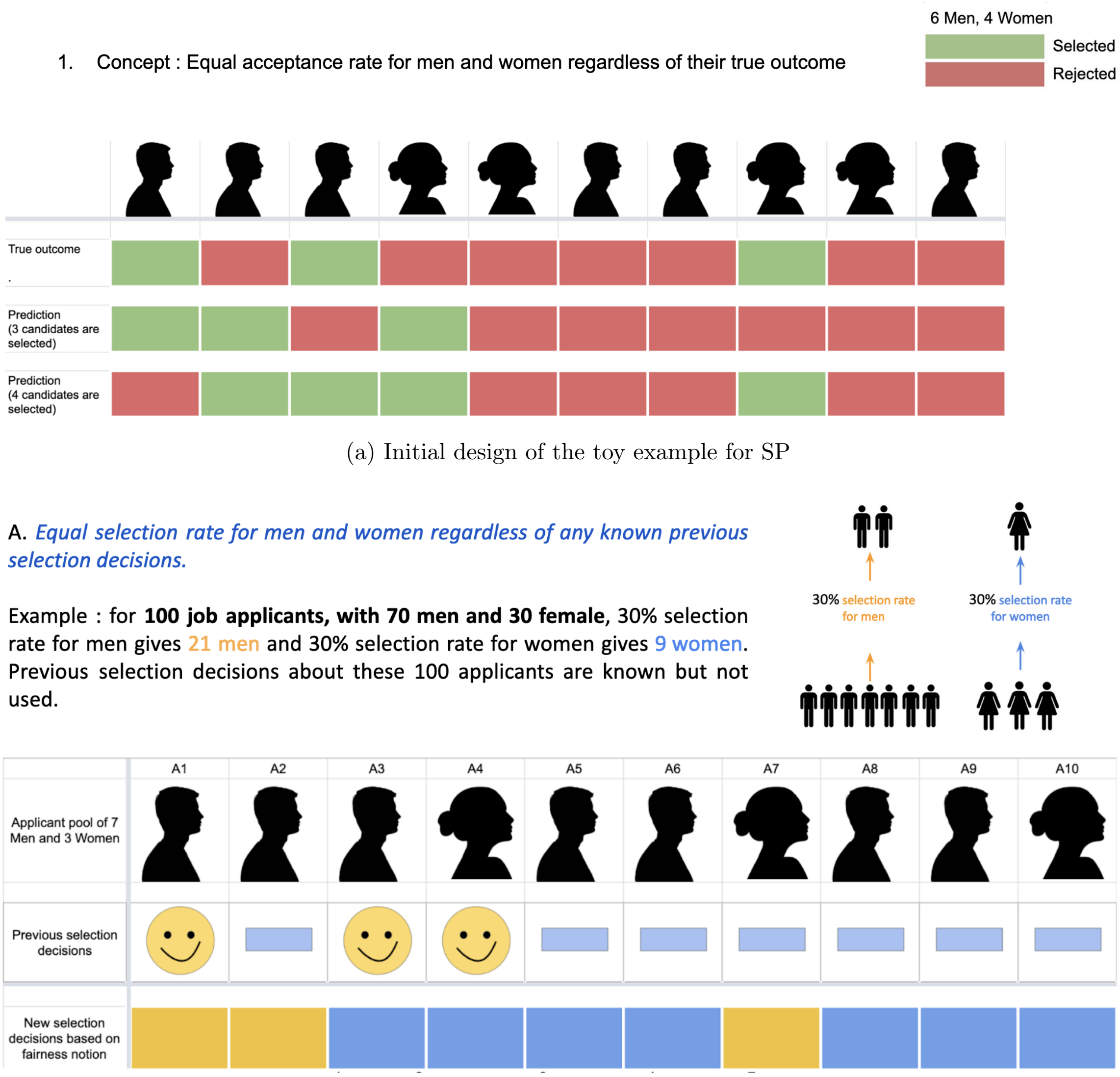
We chose 6 fairness notions for this research, from which 4 come from group fairness and the remaining 2 from individual fairness. The 4 group fairness notions are Statistical Parity (SP) [43], Equal Opportunity (EO) [59], Disparate Impact (DI) [92] and Calibration (CB) [32], as shown in Table 8. The individual fairness notions are Fairness through awareness (FA) [43] and Counterfactual fairness (CF) [80], shown in Table 7. The 6 notions are chosen based on their relevance in the context of hiring, the prevalence in available literature and as representatives of group and individual notions of fairness.

Fairness Notion	**Motivation**	**Definition**	**Useful when**	**Disadvantages**
Statistical Parity (SP)	Selection should depend only on the prediction, not the actual outcome because there is no performance data for who is not hired, and A and Y are from real world can be discriminatory	Equal prediction rates across groups regardless of the actual outcome	1) One decision is preferred over the other, 2) Legal requirement of equal acceptance rates, and 3) Dataset contains biases	1) Does not work well if unequal baserates for groups in dataset and , 2) Can be satisfied with masking (without efforts towards fairness)
Equal Opportunity (EO)	Selection should depend on the prediction for a given actual outcome because sensitive features may be correlated with prediction	Probability of being predicted in the positive class when the actual outcome is positive should not depend on the sensitive feature. Equal TPR or FPR for all sensitive groups	1) FP rate is not important, and 2) Recall is more important than precision	Not useful when ground truth is not available or dataset is biased
Calibration (CB)	Selection can be said to be fair when sensitive feature and actual outcome are not correlated	Equal probability of actually belonging to the positive class for each predicted score S=s, irrespective of the sensitive feature	Model should be independent of the decision boundary (dynamic threshold)	Difficult to achieve in practice when only a limited number of candidates can be selected
Disparate impact (DI)	Legally, different outcomes for different groups is termed discrimination and leads to impact disparity	The proportion of positive predictions from the advantaged group, say q_a and the positive predictions from the disadvantaged group, say q_b should satisfy $q_b/q_a \geq p/100$	Dataset contains biases	Can be satisfied with masking (without efforts towards fairness)

Table 8: The 4 group fairness notions - Statistical Parity [43], Equal Opportunity [59], Calibration [32] and Disparate Impact [92]

4.1.2 Mathematical definitions to user friendly explanations

The 6 fairness notions were translated into user friendly non-mathematical scenarios using an iterative process. The goal of this translation was easy and efficient communication between researchers with a background in ML and organizational representatives, who may not have a background in ML. Each fairness notion was depicted using a diagram and a textual explanation in English without using any mathematical vocabulary and discussed with a lab member with a background in social inclusion and no mathematical background in machine learning.



1. Concept : Equal acceptance rate for men and women regardless of their true outcome

6 Men, 4 Women

🟩 Selected

🟥 Rejected

True outcome	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟥	🟥
Prediction (3 candidates are selected)	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥
Prediction (4 candidates are selected)	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟥	🟥

(a) Initial design of the toy example for SP

A. *Equal selection rate for men and women regardless of any known previous selection decisions.*

Example : for **100 job applicants, with 70 men and 30 female**, 30% selection rate for men gives 21 men and 30% selection rate for women gives 9 women. Previous selection decisions about these 100 applicants are known but not used.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Applicant pool of 7 Men and 3 Women										
Previous selection decisions	😊	🟦	😊	😊	🟦	🟦	🟦	🟦	🟦	🟦
New selection decisions based on fairness notion	🟨	🟨	🟦	🟦	🟦	🟦	🟨	🟦	🟦	🟦

(b) Final design of the toy example for SP

Figure 8: Design changes to the toy example

After multiple iterations, several design changes were made shown in figure 8a for the initial design and in figure 8b for the final design.

Each fairness notion is represented using a fictitious toy example for a job application. The job type is presented as requiring merit and skill based selection, where multiple people can be selected for the position. This includes jobs such as teachers, nurses, accountants, engineers as well as students for university admission. Jobs that require specialist skills for expert positions

Fairness Notion	**User friendly definition**	**Example**
Statistical Parity (SP)	*Equal selection rate for men and women regardless of any known previous selection decisions*	For 100 job applicants, with 70 men and 30 female, 30% selection rate for men gives 21 men and 30% selection rate for women gives 9 women. Previous selection decisions about these 100 applicants are known but not used.
Equal Opportunity (EO)	*Equal selection rate for men and women from the group who was previously selected*	For 100 job applicants, with 70 men and 30 female, previous committee has selected 50 applicants of 40 men and 10 women. From this group of 50 applicants, 40% selection rate for men gives 16 men and 40 % selection rate for women gives 4 women.
Calibration (CB)	*If applicants get same points in the new selection, they should have the same possibility of previous selection or previous rejection*	We give the same points to applicants who got similar decisions before. So, we give calibrated or relative points to candidates instead of looking at each candidate individually.
Disparate impact (DI)	*Select candidates so that selection rate for women is at least 80% of the selection rate for men*	For 100 applicants, with 70 men and 30 women, select at least 4 women if 10 men are selected OR select at least 7 women, if 20 men are selected.
Fairness through awareness (FA)	*Regardless of gender, two similar applicants should be given similar decisions*	There are 2 job applicants - one man and one woman with the same characteristics except gender. We either select both or reject both because they differ only on gender.
Counterfactual fairness (CF)	*Change in only gender should not affect the selection decision*	You are given a male applicant's resume and you select him. Now, I change the gender to female on the resume while everything else remains the same. Now, you have to select her because only gender is different.

Table 9: User friendly textual definition and example for each of the chosen fairness notions

such as CEO, surgeon, Chief Scientist, Department Head are excluded in this set up. Such a set-up gives scope to include fairness and gender diversity as a strong and desirable criteria while making early stage selections. Furthermore, this set-up also allows for selection of multiple candidates for an organization.

The toy example consists of an applicant pool of 10 candidates, who have applied to a job of the description mentioned above. These 10 candidates, identified with an ID ranging from *A1* to *A10* meet all the hard requirements for the job such as education, experience and skill set, etc. This pool is composed of 7 men and 3 women, as seen in Figure **??**. It consists of a skewed gender composition to indicate that one gender applies more frequently over the years to this type of job.

For each candidate in the applicant pool, a previous selection decision is known. This is representative of the actual outcome Y or the ground truth that is available in the data to train an algorithm. A positive decision is represented by a smiley face in yellow and a negative decision is represented as a "-" symbol. Next, for each candidate, a new selection decision is also known. This new selection is representative of the prediction $\hat{Y}$ of an algorithm. These predictions are made based on one fairness notions, where yellow indicates that the candidate is selected and blue indicates that the candidate is not selected. The Figure **??** shows the toy example obtained for Statistical parity. The choice of yellow and blue to indicate selection or rejection respectively is to move away from red and green, which represents polarized decisions. Through discussions with the lab member, we concluded that red indicated that a rejected candidate should not be selected by the algorithm. The neutral colors aim to close the gap between the decisions and allow research participants to consider selections of previously rejected candidates.

Therefore, the toy example represents a binary classification problem of predicting, whether to select $\hat{Y} = 1$ or not select $\hat{Y} = 0$ an applicant. This gives a total of 6 scenarios made of the toy example, one for each of the 6 fairness notions. The Table 9 shows the fairness notion and it's final non mathematical textual explanation.

4.2 Methods

4.2.1 Recruitment of Research Participants

We used a combination of purposive sampling and snowball sampling [52] to recruit experts. Initially, we sent direct emails to relevant professionals in the domain of recruitment, Human Resource Management (HRM), Diversity and Inclusion (D&I) operations found with the help of our network of our personal connections in industry. Next, we made announcements on LinkedIn inviting professionals in the domain of recruitment, HRM, diversity and inclusion operations to participate in an interview by sending us their email. We sent follow-up emails to those who responded. We also found a list of professionals working in this field using LinkedIn, whom we contacted by sending direct messages on LinkedIn. Overall, we contacted over 80 potential participants, out of whom 48 responded. Finally, 21 participants scheduled interviews. We conducted the study in Delft University of Technology in The Netherlands.

4.2.2 Participants

The participants were interviewed between June and August of 2022, either in person or Microsoft Teams. Out of the 21 participants, we selected 17, who are most relevant to the field of this study and have experience with the selection process in the early stages of hiring. Their professional role was checked with their Linkedin profile. The demographic information of the 17 participants is listed in Table 10. The gender distribution of 17 participants were 9 (52.9%)

	Job Title	Experience	Education	Gender	Ethnic Minority*	Other minority**
E1	Talent Acquisition Specialist	23 years	Bachelor	Male	No	No
E2	D&I Officer	4 years	Doctorate	Female	No	No
E3	HR Manager	24 years	Master	Female	No	No
E4	Managing Director	3 years	Master	Male	No	No
E5	Executive Board Member	5 years	Bachelor	Female	No	Yes
E6	CEO	14 years	Master	Male	Yes	No
E7	HR Development Trainee	1 year	Master	Male	No	No
E8	HR Business Partner	4 year	Master	Male	Yes	No
E9	HR advisor	5 years	Doctorate	Female	No	No
E10	D&I Advisor	8 years	Master	Female	No	Yes
E11	Chief Diversity Officer	10 years	Doctorate	Male	Yes	No
E12	Recruitement Technology Consultant	15 years	Master	Male	No	No
E13	Assistant Professor (as vacancy holder)	6 years	Doctorate	Female	No	No
E14	Psychological Assessment Reseacher	40 years	Master	Female	No	No
E15	Global D&I Manager	12 years	Master	Female	Yes	Yes
E16	Inclusion Specialist	6 years	Bachelor	Male	No	Yes
E17	I/O Psychologist	7 years	Master	Female	Yes	No

Ethnic Minority* = self-reported as belonging to an ethnic minority
Other Minority** = self-reported as belonging to a minority group that faces discrimination

Table 10: Demographic Information of Study Participants

Employment Industry	Number of Participants	Organization Size
Education and Sports	1	<10
Recruitment Service	2	<10
Water Management	1	<100
Human Rights	1	<100
Arts and Culture	1	<100
Higher Education and Research	3	<1,000
Higher Education and Research	4	<10,000
FMCG	1	<70,000
Police	2	<70,000
Offshoring	1	<70,000

Table 11: The number of participants per employment industry and organization size

female and 8 (47.1%) male, with an average of 11.6 years (SD=9.67) of relevant HR experience. 10 (58.82%) participants have a Master degree, 4 (23.53%) participants have a Doctorate degree and 3 (17.65%) participants have a Bachelor degree. Out of the 17 participants, 5 (29.41%) self-reported as belonging to an ethnic minority, 4 (23.53%) self-reported as belonging to other minority related to age, health, sexual orientation, immigration and neurodiversity, and only 1 (5.88%) belonged to both ethnic and another minority group. All participants work in The Netherlands. Their current employment industry, along with the size of the organization is shown in table 11.

4.2.3 Interview Process

The same researcher interviewed each participant independently, either in person or online using Microsoft Teams. The semi-structured interviews were conducted in one or two sittings session and lasted 75 minutes on average. The interview began by presenting an opening statement containing goal of the interview, participant information to be collected, purpose of audio recording, participant's right to withdraw from the research and contact details of the researchers. After obtaining the participant's explicit and voluntary consent to the opening statement, we proceeded with the semi-structured interviews.

The interview consisted of 5 parts as described below:

1. Introductions were made and the participants provided their background information, shown in Table 10.
2. The participants were introduced to the context of early candidate selection, example job positions and multiple vacancies. Further, the participants were not given any algorithmic context - actual outcome was represented as decisions made by a committee, whose 'yes' and 'no' decisions were known but not their criteria for selection. The origin of the fairness notions was not disclosed to the participants. The deliberate decision to not disclose algorithmic context was taken to get a bias free commentary from the participants because the type of decision maker for the same task evokes different emotions in people [84].
3. Participants were shown a user friendly definition of one fairness notion and asked to think out loud about their understanding, concerns, benefits, feelings and implications of adopting the fairness notion. After participants described their initial thoughts, they were shown the pictorial representation of the same fairness notion and their comments were obtained. The discussion of each fairness notion concluded with their rating on understanding, fairness and diversity in terms of gender, which is described in the following questions:

    - *What would you rate your understanding of this fairness notion?*
      (1=Don't Understand, 2=Somewhat Don't Understand, 3=Don't know, 4=Somewhat Understand, 5=Understand)
    - *What would you rate this notion on fairness?*
      (1=Unfair, 2=Somewhat Unfair , 3=Don't know, 4=Somewhat Fair, 5=Fair)
    - *What would you rate this notion's ability to improve gender diversity?*
      (1=Unhelpful, 2=Somewhat Unhelpful , 3=Don't know, 4=Somewhat Helpful, 5=Helpful)

    These rating at the end of each fairness notion was collected on a 5 point scale where, 1 indicates the lowest value and 5 indicates the highest value. The process was repeated for 6 fairness definitions. Participants were shown two individual fairness notions and 4 group fairness notions in random order to mitigate order effect. However, the order within 4 group fairness was in the following order: SP, EO, DI and CB and the order within 2 individual fairness notions was kept to FA and CF. This was done to help the participant move from the simplest to hardest user friendly translation as some learning curve was observed in the pilot studies.
4. The participants were asked to rate each fairness notion again using the same 5 point scale. This was done to handle order effects in the previous part.
5. Participants were asked to describe their experience of the interview and their thoughts about using the fairness notions.

4.2.4 Analysis

We used thematic analysis [21] to code approximately 26 hours of audio or video recording and their corresponding transcripts using Atlas.ti. The transcripts were obtained from closed captioning functionality of Adobe Premiere Pro 2022 for audio recordings and Microsoft Teams for video recordings, which were all checked for correctness of translation. Initially, two of the researchers worked on open coding 4 interview transcripts using inductive approach, which was later compared and consolidated to establish coding practice and guidelines for consistency in the research. Next, all the 17 transcripts were coded using deductive approach to answer the research questions. After deliberation and discussion, the codes were categorized and formulated into high level themes.

4.2.5 Ethical Consideration

The research was approved by Human Research Ethics Approval of Delft University of Technology, which consisted of a Data Management Plan, Risk Evaluation Checklist and Opening statement as Informed Consent for the research Participants.

5 RESULTS

This section shows the results obtained from the analysis described in section 4. We present the ratings provided by participants for three factors Understanding , Fairness and Diversity of using the mathematical fairness notions on a 5-point Likert Scale. 17 participants rated SP, EO, CB, FA, CF and 8 of them also rated DI as shown in Figures 9 through 14. Each figure shows the number of participants and their corresponding ratings for the three factors. The ratings are further broken down by sensitive features - *Gender, Ethnic Minority* and *Other Minority*, to examine potential relationship between a participant's sensitive feature and their ratings. *Gender*, as reported by participants contains two values - Male (M) and Female (F). *Ethnic Minority and Other Minority* is represented as a *Yes* or *No*.

Next, we qualitatively discuss the four core themes - impression, perception of fairness, perception on improving diversity, and applicability - obtained from the thematic analysis performed on all interview transcripts. Lastly, we present the ratings on fairness again, given by participants after reflecting on all the fairness notions. We conclude this section with a discussion of the changes in their ratings.

5.1 Statistical Parity (SP)

We present the qualitative analysis for SP from 150 quotations obtained upon coding transcripts of 17 participants. All the ratings given by participants for three factors - Understanding (U), Fairness (F) and Diversity (D) of using SP are shown in Figure 9.

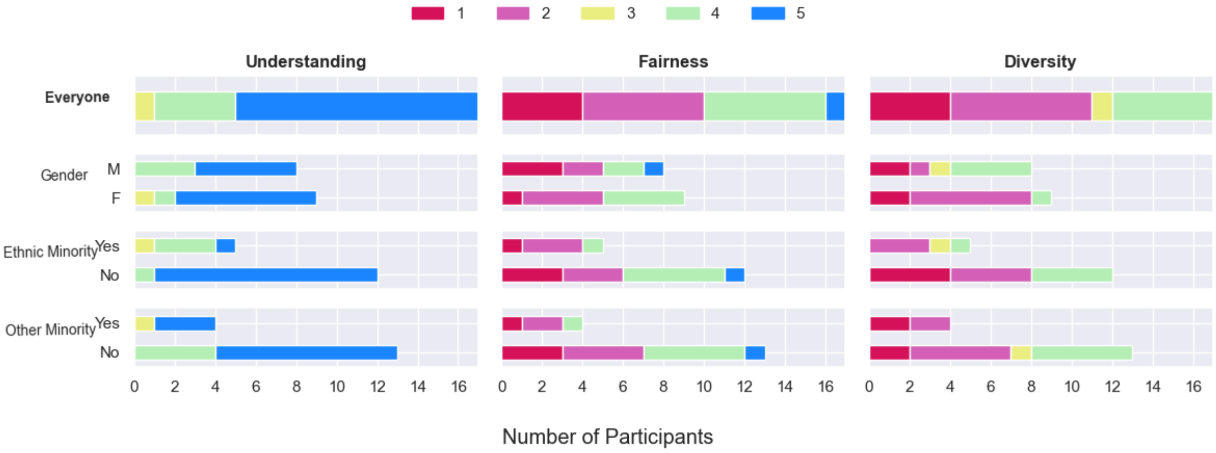


Figure 9: **SP** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of SP, along with the distribution by *gender, ethnic minority* and *other minority* groups.

5.1.1 What is their impression of SP?

Participants had a wide range of thoughts and comments on the notion of Statistical Parity. Six participants expressed what they liked such as merit being the basis of selection (E1, E9, E12) and that the notion objectively treats all job applicants equally (E2, E4, E6). *"..it should not restrict us getting in contact with good candidates"* - E9. *"creates equal opportunities for both sexes, which is positive"* - E6. However, two of the same and four additional participants also disagreed with the notion or said that they would not use the notion in their organizations (E5, E6, E8, E11, E12, E16). *"I understand what it's trying to do. But would I use it? Do I agree with the fairness notion, then no."* - E5. *"I would never do it this way. I don't care what the* *gender is in this stage. I want to select the best candidates."* - E12. In fact, several participants expressed their dislike for the notion directly or indirectly. While some participants directly mentioned that they did not like the notion (E5, E6, E7), others did not like gender being the basis of separation (E4, E1, E7). *"I don't like this one. It is very well possible that these 70 men aren't very qualified at all and then you're going to still hire 30% of them"* - E7. *" ..you are making a difference on gender and it's something we try to avoid as long as possible."* - E4. Majority of the participants understood the notion and gave a rating of 4 or 5.

While one participant was glad that Statistical Parity allowed some representation of minority groups, which might not otherwise happen without effort (E10), three participants said that the notion favors majority and may not help minorities (E2, E10, E15). *"I'm happy that there's at least more than just one woman, because we often see there's only one woman. Looking at chances, the men have more chance of being hired than the women do in this case."* - E10. *"We know that the chance of hiring female is zero. Right?" (sighs)* - E11.

5.1.2 What is their perception on fairness of using SP?

There was no consensus in the ratings, with four participants rating it 1 and an equal number of participants rating it 2 and 4. A similar pattern of ratings is seen for both genders and majority groups. However, *ethnic* and *other minority* provide a low rating on fairness. Thematic analysis of their responses explains the ratings.

About nine participants, argued this notion as unfair because they feel it is not equitable (E8), is based on gender (E5, E6), is based on quota system (E11), skewed applicant pool (E3, E15, E16) and other factors (E1, E2). *"it's not equitable. So I think it's unfair"*. - E8. *"The whole system is not fair because you don't want to use a quota system, right?"* - E11. On further examination, we see that many participants consider Statistical Parity as only theoretically fair and applicable only in an ideal world (E2, E4, E8, E9, E13, E15, E16, E17). *"It's fair in theory, unfair in result. If it were a fair world, a perfect world, then this would be a fair procedure."* - E2. According to them, theoretical fairness is attributed to ideally treating everybody equally. *"Technically, it's still fair because of same selection rate for men and women"* - E4. However, seven of them hesitated about the calling the notion fair. *"I think it's fair on paper [...] it's actually not quite fair because 30 women and 70 men applied for the job, and it feels really wrong to have only 9 women go on to the next round"* - E16. Lastly, four participants found it quite difficult to say whether or not this notion could be called fair. *" this is a difficult one. I really would like to know why A1, A3 and A4 were chosen by the former committee and why the decision is now different. If I don't know why, I cannot say if it's fair or not, or sensible."* - E14.

5.1.3 What is their perception on improving diversity on using SP?

Participants were more certain about commenting on improving gender balance in organizations with the help of Statistical Parity. Five participants felt positive about the notion's ability to help diversity in terms of gender (E1, E2, E4, E7, E9). *"We have 75% men and 25% women in our company. So when we introduce this fairness notion, then it will help to balance our company more."* - E4. They like the notion and reasoned that despite a skewed applicant pool it can act as a precautionary step because final hiring can be biased, however long the improvement takes. *"You will provide possibility of balance in your selection group but then there's still the final selection"*. - E1. *"But if you really want to hit the targets, that's probably not going to help."* - E7. This can be seen in the ratings, where seven participants gave a rating of 4 or higher for diversity.

However, a larger number, totaling eight participants, reasoned about the notion's inability to improve diversity for two main reasons(E2, E3, E5, E8, E12, E13, E15, E16). First, four participants said that diversity goals cannot be achieved if minority is absent in the applicant pool. *"Because if only one female applies, then there goes your theory"* - E5. Second, six of them said that broader diversity goals of the organization cannot be realized because of small minority representation in selections. *"You need to identify your minority group and make sure that's at least 50% of your talent pool. Otherwise you'll never make this change."* - E8. Despite, polarized perceptions, five participants indicated that the notion could contribute to diversity to some extent (E1, E4, E6, E11, E13). *"I mean of course, if your team had no women then you're improving your gender balance. If your team consisted of only women, you might want to select only men. So it depends on what your gender balance was."* - E13. From the ratings, we see that 10 participants rated 1 or 2 for diversity.

5.1.4 What is the applicability of SP?

Majority of participants were concerned about the structure of the selection rate. Eight of them said that the status quo of minority and majority would not change resulting in unfairness towards minority (E5, E8, E9, E10, E11, 15, E16, E17). *" I presume that you want equal representation and you can't do that by focusing on percentages because as you see the end of the funnel, you'll still end up with a majority and a minority."* - E5. Five of them were concerned about gender, instead of merit being the focus of selection (E1, E4, E5, E7, E12). *"You're not looking at the big picture of hiring the best candidates. This is statistics."* - E12. *"I hope that most companies won't make the decision focused solely on gender."* - E5. Lastly, Four of them said that equal selection rate removes effort towards fairness or attracting more diverse applicants (E2, E10, E12, E15). *"Apparently [here], we see men as higher quality than women and the risk is that you have this excuse woman. We need a woman in selection procedure. We we have more men, but we also have one woman. So we are also diverse"* - E10. *"I see that we have overwhelmingly male applicants or female applicants. Whatever the role is, I'm always curious why that is the gender distribution. Is it something about our job ads? Is it something about the language that we use?"* - E15. Interestingly, some of them suggest that Statistical Parity is applicable only when the applicant pool is large (E5, E2) and contains only qualified (E1, E7)and diverse (E15, E16) applicants, which would increase it's effectiveness (E2, E15). *"When you create models like this, you often take the presumption that many will apply. But what will happen if you only have 3 applicants?"* - E5. *"If all the hard criteria is met, the percentages would make more sense and then they would feel more fair."* - E1. Furthermore, many participants also suggested modifying the selection rate by making it proportional to the applicant pool representation (E5, E10, E16), making the selection rate higher for minorities (E2, E15, E16, E17), or opting for minimum number of minority candidates over a percentage (E11). *"It doesn't feel fair looking at it from this perspective. I feel like the selection rate might need to be higher for women because there are fewer women if they're all qualified."* - E15. *"There needs to be at least one female candidate on the shortlist. It is kind of a minimum requirement of one at least one viable female candidate. Otherwise you have to keep searching. You can't only have male candidates."* - E11.

Another major source of concern among the participants was the context in which the notion is applied. *"It's a clear notion, but it misses lots of context."* - E6. They said that it's applicability depends on the the composition of the existing team and type of job role, saying that a more diverse employee base and a generalist role would make Statistical Parity fair to use (E1, E5, E6, E13, E17). *"I would also argue that it depends on your current team if you want to have a diverse team. If you have a team of only women, then you could argue to put more focus to men."* - E6. *"If you copy this model to a production facility with 150 people doing almost exact* *the same work, then it would be easy, really easy to implement"* - E1. Some of the participants also pointed that the notion's applicability depends on the type of organization, its size and more importantly its goal (E5, E6). *"You can't just replicate it towards an entire industry or even jump function or organization. It wouldn't create the effect you're looking for I think. In SMEs talent pools aren't that big."* - E5.

5.2 Equal Opportunity (EO)

All the 17 participants discussed EO after SP because SP could help bridge the learning gap needed to reflect on EO. We obtained 88 quotations upon coding transcripts of the 17 participants. All the ratings given by participants for three factors - Understanding (U), Fairness (F) and Diversity (D) of using EO are shown in Figure 10.

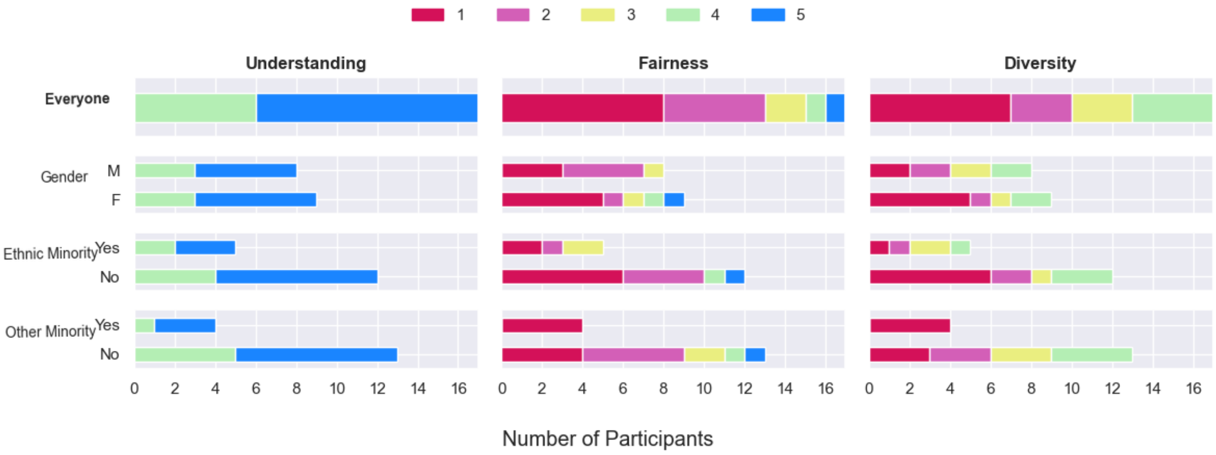


Figure 10: **EO** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of EO, along with the distribution by *gender, ethnic minority* and *other minority* groups.

5.2.1 What is their impression of EO?

Many participants, having talked about Statistical Parity were able to point the similar underlying principle in Equal Opportunity (E5, E9, E12), where some even felt the current notion being less fair (E1, E2, E4, E7). *"It's slightly different data, but the principle is the same."* - E12. *"I don't like it even more than the previous one because of this." (points to actual outcome)* - E2. All participants gave a rating of 4 or 5 for their understanding of EO. While all the participants understood the notion, some with questions, about five participants immediately indicated their dislike or disagreement with the notion (E8, E10, E11, E12, E17). *"I would never use this. It's a forced way [what] you're doing. But this is not what helps you realize non biased selection."* - E12. *"I don't agree. The problem is I don't agree with with selecting people on the basis of splitting up on percentages and gender. I think it will create conflict and upset people even more because it's just based upon numbers."* - E11. However, a large source of dislike came from the use of actual outcome in final predictions. Nine participants either asked for reasons for the actual outcome or outright disagreed with its usage saying that it can influence a human selection maker if actual outcome is known (E1, E2, E4, E5, E7, E10, E13, E16, E17). *"It will it will interfere with your selection method knowing what somebody else concluded."* - E1. *"If you ask me what happened and and why is this like this, why did they do this then? That part I don't understand."* - E17. It is important to point out that the actual data can only be used while training the model, and may not be shown to the human decision maker. The participants, however were not aware that the prediction $\hat{Y}$ is made by the model and not a human.

5.2.2 What is their perception on fairness of using EO ?

Majority, totaling 13 participants, rated the notion either a 1 or 2 on fairness. Four participants said that Equal Opportunity is less fair than Statistical Parity (E1, E2, E4, E7). *"If you do not know the background of the decisions, I would still go for an equal selection rate of the applicant pool [SP]"* - E4. While only one participant indicated the theoretical fairness of EO (E16), about seven participants expressed the notion being unfair due to the influence of actual outcome in the selections and the need more information on reasons for actual outcome (E1, E2, E4, E7, E10, E15, E17). *"Has to be very clear on what basis those decisions are being made. And in this scenario, it wasn't clear. So that's not fair."* - E7. *"What was the selection criteria and if that's unknown to me, then this new selection doesn't seem fair because I don't have the information to make that decision"* - E15. This reason also made three participants unwilling to provide a rating on fairness (E7, E14, E17). *"I have no clue. I couldn't also not say if it's fair or not."* - E17. *"For now we are going to do a equal percentages. But when your organization needs more women or maybe more men, then you have to have other principles. But for now I cannot say anything else."* - E14.

5.2.3 What is their perception on improving diversity on using EO?

A majority, totaling 10 participants rated 1 or 2 on the ability of EO to improve diversity. Similar patterns can be seen across genders and minorities, with no participant giving a rating of 5. From the thematic analysis of their responses we see that many participants were certain that using EO will not improve gender balance in organizations. Six of them attributed this to small minority in the applicant pool (E1, E4, E7), their slim chances of being selected in the actual outcome (E4, E11) and lack of trust in the actual outcome (E1, E13, E15). *"Then you most probably end up with three or four men and zero or one women."* - E4. *"You're copying the same bias, perhaps as the previous person"* - E1. The remaining participants did not provide reasons for their ratings on diversity.

5.2.4 What is the applicability of EO?

The biggest source of concern for majority of the participants was trusting the actual outcome. Six participants mentioned that bias present in the actual outcome will get copied to the prediction (E1, E2, E5, E10, E13, E16) defeating the purpose of the fairness notion. *" If this (points to actual outcome) is very unfair, then it propagates unfairness."* - E2. *"What we do as an organization is that we never ask the opinion of the previous committee."* - E10. While, 11 participants expressed concerns about using the actual outcome, three participants mentioned that they would only use the notions if the reasons for actual outcome is known (E6, E7, E9). *"Knowing what somebody else concluded will interfere with your selection method."* - E1. *"The first thing that I will do is check with them. On what basis did you select those people?"* - E7. *"Sometimes you would have to trust that people made the right decisions and you have to move from there."* - E9. One participant also said that usability would improve if there was space to disagree with the actual outcome (E16). *"Instead of A1, I think I would like to interview A7 or A10 only because I'm very curious to see what they're about and just to find out if there is anything that the previous selection decisions [missed]."* - E16.

The next set of concerns affecting the applicability of EO was its inability to help diversity goals. Participants said that diversity goals could not be achieved using equal selection rate and gender, in the example was used just for sake of diversity, thereby decreasing effort towards fairness (E9, E11, E12). *"This equal selection rate for merit should be something to monitor but not to aim for."* - E9. *"You're confusing the situation, you're confusing things by doing this. You're just using gender to make selections without any reasoning. You can play with the percentages, you can create all kinds of different equations, but it doesn't serve justice to what you want to achieve in the end, right?"* - E11. Four participants felt that just like Statistical Parity, the applicability of Equal Opportunity was affected by a skewed applicant pool, which favors majority, suggesting a higher selection rate for minorities (E5, E7, E8, E10). *"If the basis you're working from is not truly inclusive, you'll see that with every cycle that difference and imbalance magnifies."* - E5. *"If they're all qualified, if they can all do the job, why not select women? Because it would be very good to restore gender balance. In the long term, it's almost always better for your company."* - E7. Lastly, one participant said that diversity goals can be achieved by involving multiple stakeholders, which EO currently misses (E10). *"You really need these different perspectives, put them together and then you can, I think create fair principles and fair ways of working. This [EO] was probably made by one person."* - E10.

5.3 Calibration (CB)

We obtained 72 quotations upon coding transcripts of the 17 participants. All the ratings given by participants for three factors - Understanding (U), Fairness (F) and Diversity (D) of using CB are shown in Figure 11.

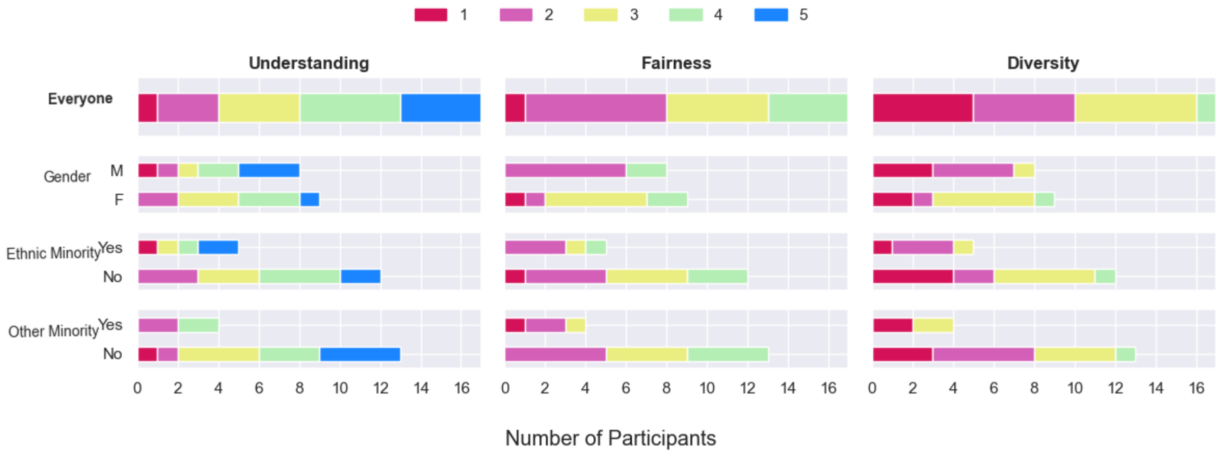


Figure 11: **CB** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of CB, along with the distribution by *gender, ethnic minority* and *other minority* groups.

5.3.1 What is their impression of CB?

About half of the participants rated their understanding of CB at 3 or lower. Five additional participants rated it 4 and only four participants gave a rating of 5. Most participants said that they found it confusing or illogical (E1, E4, E5, E6, E7, E9, E10, E13, E15, E16, E17). *"I'm trying to find the logic behind this. I can can really find it"* - E17. *"I think this system is really confusing. I think I don't see why this would help"* - E16. They said that it felt mathematical or asked clarifying questions on the scores (E1, E9). *"It looks very scientific and quantitative. But I know from practice is that can be very difficult"* - E9. *"The points are based on what exactly?"* - E17. Despite the difficulty in understanding the concept, three participants pointed out rating participants on merit is fair and can help keep human biases in check (E1, E4, E11). *"If you purely look at merit based recruiting, grading candidates based on* *their merits is fair."* - E1. Three participants extended the discussion saying they liked that applicants were compared, which can remove the myth of meritocracy (E1, E2, E17). *"I would say the positive will eliminate this meritocracy myth."* - E2. However, two people also disliked comparing applicants and assigning them scores (E4, E10).

5.3.2 What is their perception on fairness and diversity of using CB?

About 13 participants rated fairness of CB as 3 or lower. However, not all participants were able to clearly express reasons behind their ratings. Two participants indicated that they needed more information and context to judge CB on fairness (E13, E17). *"I can't really say anything about fairness because I have no clue about the previous decisions. I really need that piece of information to. Make a statement about fairness."* - E17. Three other participants felt that the process of assigning scores to applicants was unclear and they could not justify (E4, E5, E6). *"The idea of grades can help explain to people why they were selected. But in this situation, it's quite difficult because if you want to be transparent and open on it, I would say this is really difficult."* - E4. *"I don't understand how you can give points in this way to candidates. That's why it feels not fair if it's not transparent enough and how the points are made and how decisions are made. It lacks transparency."* - E6.

5.3.3 What is their perception on improving diversity of using CB?

Majority of the participants rated 3 or lower for diversity. While 10 participants rated it 1 or 2, a stark number of 6 participants rated it 3, indicating that they didn't know if CB could help diversity in organizations. Eight participants provided reasons for their ratings on diversity saying that they were either uncertain or could not see how gender balance would improve with the help of CB (E1, E2, E4, E5, E6, E9, E10, E17). *"If you crystallize the process, then could it help on gender balance? Maybe. I wouldn't dare say at this point."* - E5. *"I think this has nothing to do with gender balance"* - E10.

5.3.4 What is the applicability of CB?

Participants expressed two major concerns regarding the applicability of CB. First, 8 participants said that the process of assigning scores was unclear making it lack transparency and logic (E2, E3, E7, E10, E11, E13, E14, E16). *"It looks like it's very objective, but it's just a number and you don't know what it's based on"* - E3. *"That makes no sense. And I don't think it would contribute to anything. I don't think it would help gender balance. I don't think it's logical."* - E11. Further, they expressed that lack of transparency made CB undesirable for use (E2, E4, E6, E11, E14). *"What actually do you take into account whether something can be quantifiable, which is known to be more in favor for men like publications or grants? Then you are by default lowering the values of women."* - E2. *" If I don't understand it, I can't see how it's going to help me."* - E6. The second major concern came from doubts about actual outcome. Eight participants expressed theirs doubts about using it and said that it was irrelevant to use it in decisions (E2, E7, E10, E11, E12, E14, E14, E5, E7). *"It can propagate bias in selection, right?"* - E2. *"I don't know why they made these decisions. I don't know what the selection committee was like. I have no idea. So that makes me not like using any of their previous decisions."* - E15.

5.4 Fairness Through Awareness (FA)

Half of the participants were shown individual fairness notions before group fairness notions. FA was the first out of the two individual fairness notions, which means that half of the participants started the interview with FA. We obtained 127 quotations upon coding transcripts of the 17 participants. All the ratings given by participants for three factors - Understanding (U), Fairness (F) and Diversity (D) of using FA are shown in Figure 12.

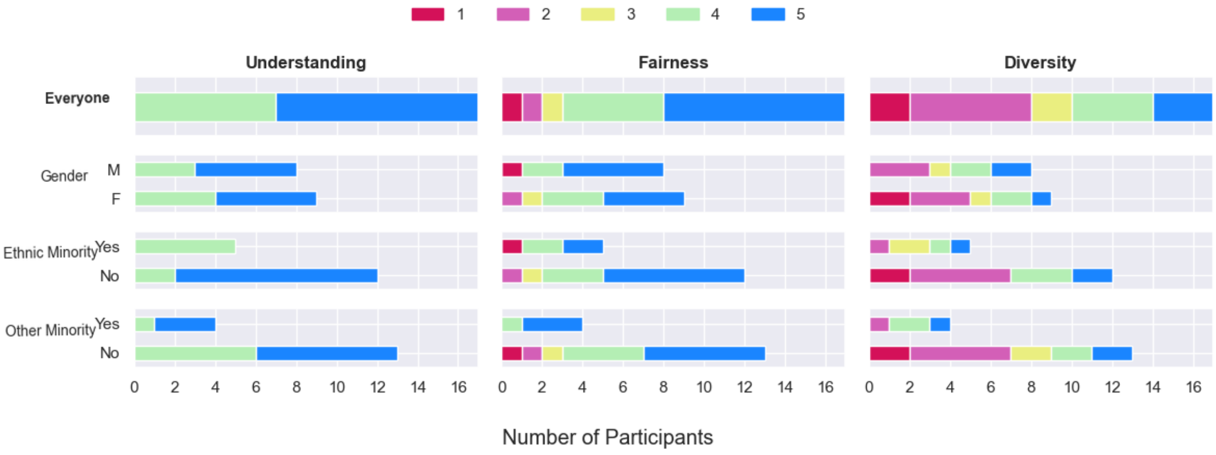


Figure 12: **FA** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of FA, along with the distribution by *gender, ethnic minority* and *other minority* groups.

5.4.1 What is their impression of FA?

Participants had many things to say about FA. While six participants agreed with the notion saying they were happy that both genders were treated equally (E2, E4, E5, E6, E9, E17), two participants found it difficult to comment on the notion (E1, E9). *"I agree. And I would select both."* - E4. *"We cannot disagree with that. That is what we should aim for"* - E9. On further probing 8 participants mentioned that in their experience two similar people are never the same, saying that they may differ on some aspects such as potential or soft skills (E1, E3, E4, E5, E8, E9, E13, E15). *"They're all unique, so it might differ in terms of location, match or growth in different direction"* - E5. *"So they had their PhD, say in 2012 and they both have, I don't know, 16 papers. But if the woman has been on maternity leave twice, then those 16 papers means she's been in the other time much more productive than the man. So what do you even mean with the same characteristics?"* - E13.

Every participant rated their understanding at 4 or 5, indicating that they understood the definition of FA. While, majority of the participants clearly said that they understood the notion, most of them also indicated that this notion reflects the ideal end goal cautioning its use to only early phases of hiring (E2, E5, E8, E9, E13, E14, E16). *"I understand it fully. But circumstances make it sometimes impossible to follow this principle."* - E14. *"I think it might work best in the beginning of the hiring process. Because, you know, in at the end of the process, you might make decisions on different data points"* - E5. *"I think it's where you want to go as an end goal but to drive the change for unconscious bias, need to create more opportunities for women"* - E8.

5.4.2 What is their perception on fairness of using FA?

Majority of the participants rated fairness of FA at 4 or higher. Interestingly, they reflected on many dimensions to assess the fairness of FA. While four participants referred to FA as being principally fair (E6, E8, E10, E11), several others provided reasons for calling the notion unfair. *"The principle in itself is rationally fair. The world that we live in is not."* - E10. Participants called FA unfair because it is not equitable, cannot achieve diversity goals, disadvantage to minorities (E2, E7, E8, E15). *"I think given the gender imbalance for which we want to correct, then I don't think this is fair "* - E2. *"This is a utopia idea. I think it's still inequitable because the female will always have a bias, that she is considered less fit for the job if everything else is the same"* - E8. Further, participants reflected that while FA might be fair towards applicants (E1, E3, E5, E7, E12, E16, E17), minorities groups might still find it unfair (E1, E14, E15) and they felt that it was more important to fix historical justice because majorities won't be severely affected if minorities are given more opportunities (E2, E8, E11, E15, E17). *"It's difficult for the male candidate, but we have a kind of historical injustice that has to be fixed. It doesn't mean we only select female applicants, right?"* - E11. Lastly, four participants found it quite difficult to rate fairness of FA saying that fairness depends on many factors such (E9, E10, E11, E13). *"I find it very hard to answer because I need a more clear definition of what the the same characteristics are."* - E13. *"It's fair within the scope of the context, which I would define but as a general statement, it's not fair."* - E11.

5.4.3 What is their perception on improving diversity on using FA?

In contrast to ratings on fairness, ratings on diversity are however, divided across the Likert scale and there is no consensus on it, with the mean at 3. While three participants said that FA could help organizations improve gender balance (E4, E5, E15), seven other participants said that FA could not improve gender balance because it is prone to cultural cloning, skewed applicant pool and minorities self-selecting themselves for job applications (E2, E6, E8, E9, E10, E14, E16). *"If you have a black woman and you have a white man and the previous 20 people that did the job were all white men, then the white man will be hired again. This is how our brain works."* - E10. *"If you want to improve gender balance in an organization, then sometimes you cannot have this fair criterion. Sometimes you have to let that go."* - E9. Lastly, five participants indicated that the notion could help in some contexts, while being supported by other measures within the organization (E3, E4, E7, E11, E17). *"Yes, it will help improvement, but it's not the only thing."* - E11. *"It depends on so many things. I don't know how many women or men employees the organization has."* - E17.

5.4.4 What is the applicability of FA?

Majority of the participants said that FA is quite theoretical and difficult to apply in practice (E1, E2, E5, E8, E9, E10, E11, E13). *"It's the most fair thing to do, but it doesn't make any sense"* - E1. *"In the perfect world where everybody would be treated equally and we have equal opportunities, then this principle is great but we are not in that world right now."* - E10. *"On paper, it may look very easy, but in practice it's not."* - E9. Additionally, five participants also indicated that the notion benefits the majorities more than it benefits the minorities and cannot select applicants who have different characteristics than previous employees (E1, E3, E4, E8, E10). *"It could be that a male applicant has more profit of this fairness notion than a woman because maybe a woman was pregnant a couple of times"* - E3. *"So if you have always had white men of a certain age with a certain background, certain studies, and they were always doing the job in a good way, you will pick a person that is the same as all the people that did it before"* - E10. Lastly, some participants mentioned that the notion would be applicable only in organizations that are diverse (E7, E10, E11, E14). *"When [your organization] is balanced and you are in that sense equal, you're also giving equal opportunities to everybody. Then you can apply this principle of fairness for sure."* - E10.

5.5 Counterfactual Fairness (CF)

All the 17 participants discussed CF after FA because FA could help bridge the learning gap needed to reflect on CF. We obtained 72 quotations upon coding transcripts of the 17 participants. All the ratings given by participants for three factors - Understanding (U), Fairness (F) and Diversity (D) of using EO are shown in Figure 13.

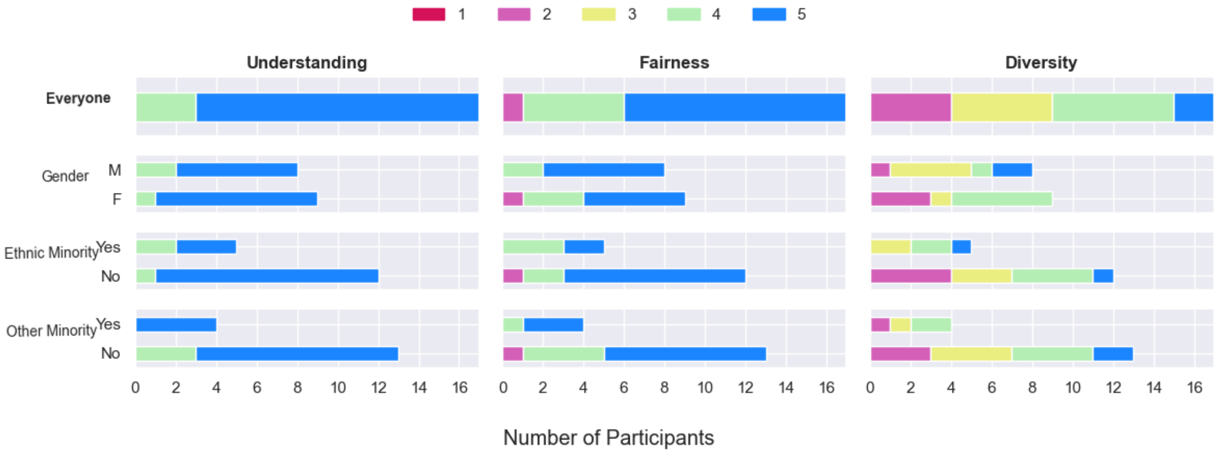


Figure 13: **CF** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of CF, along with the distribution by *gender, ethnic minority* and *other minority* groups.

5.5.1 What is their impression of CF?

Every participant rated their understanding at 4 or 5, with majority of the participants rating it 5. Majority of the participants either understood or agreed with the notion (E1, E3, E6, E7, E8, E9, E10, E12, E13, E14), and two of them expressed curiosity about trying out the notion (E1, E16). *"I agree that changing only gender should not affect the selection decision."* - E14. *"I really like this especially if you test it."* - E1. Additionally, some participants mentioned that such a notion can help keep discrimination in check (E1, E2, E16). *"You're not pushing more women forwards or more men forwards. You're just checking bias"* - E1. Lastly, few participants indicated the similarity of CF and FA (E1, E7) *"I think there's they're ethically speaking the same, right?"* - E7.

5.5.2 What is their perception on fairness of using CF?

Perception of fairness and diversity of using CF elicited few responses from the participants. Five participants said that the notion is fair in principle but is not very practical or useful (E5, E10, E10, E14, E17). *"The principle in itself is very fair. But putting it in the context of our world becomes very complex."* - E10. Another set of four participants felt the notion is fair because it is purely based on merit and could help improve the selection of minorities (E7, E9, E11, E15). *"I would rate this as fair because nothing has changed about the qualifications of the candidate at this point."* - E15.

5.5.3 What is their perception on improving diversity of using CF?

Ratings are divided for diversity across the Likert scale and there is no consensus on it. Majority of the participants were unsure whether CF would help improve the gender balance in organizations and indicated the using other measures instead (E2, E3, E4, E5, E6, E8, E9, E10, E12, E14). *“Can it help improve gender balance? I’m not sure. This notion, not by itself, no."* - E6. *“No, it won’t help. When you want to have more women in your organization, Then sometimes you have to [use other approaches]"* - E14. *“I don’t think in in terms of fairness or diversity it can be improved. At least temporarily, we might need different measures."* - E5. Few participants indicated that depending on the context, it could help diversity (E11, E16). *“We want to get more males for secretarial support. There, it would contribute to helping the gender balance."* - E11.

5.5.4 What is the applicability of CF?

The biggest concern about CF expressed by participants was related to gender. Four participants said that they were unsure about gender or another sensitive feature being part of the selection process when CF is used (E1, E6, E9, E17). *“I think it’s all information that we do not need for open, transparent and merit based recruiting. It might help with diversity, though. And sometimes a big age gap can be seen as something less positive too"* - E1. While, three other participants said that diversity is a quality and such features can be helpful in selection (E3, E4, E10). *“To me and to our organization, diversity is also a quality. So, if the norm in your organization is female and the applicant is male, then that is also a quality of the person"* - E10. *“Men and women bring different perspectives which you should also consider. You don’t see this if you remove the gender."* - E3. Two additional participants indicated low applicability of CF, if diversity is a goal of the company (E7, E11).

On the applicability of CF, four participants said that improvement would be quite slow because the notion favors majority (E2, E5, E13, E15). *“You have marginalized groups that are like 10 - 0 behind. So you can use notion, but that would mean that we would have to go to many cycles to reach to reach true fairness levels. So it’s not always applicable."* - E5. Lastly, two participants indicated that the notion is not bidirectional, meaning that it was not logical to show majority as a minority applicant because they would still be favored (E2, E13). *“I think what occurs more is that you would have females who are not selected and then if they were presented as a male candidates, they would be selected and not the other way round."* - E2. *“You can never have the same candidate’s gender flipped because a man cannot become pregnant, so you cannot have a gender neutral CV because how do you account for pregnancy leave then."* - E13.

5.6 Disparate Impact (DI)

Eight out of the 17 participants were interviewed on DI. We present the analysis for DI from 30 quotations obtained from corresponding transcripts. All the ratings given by participants for three factors - Understanding (U), Fairness (F) and Diversity (D) of using DI are shown in Figure 14.

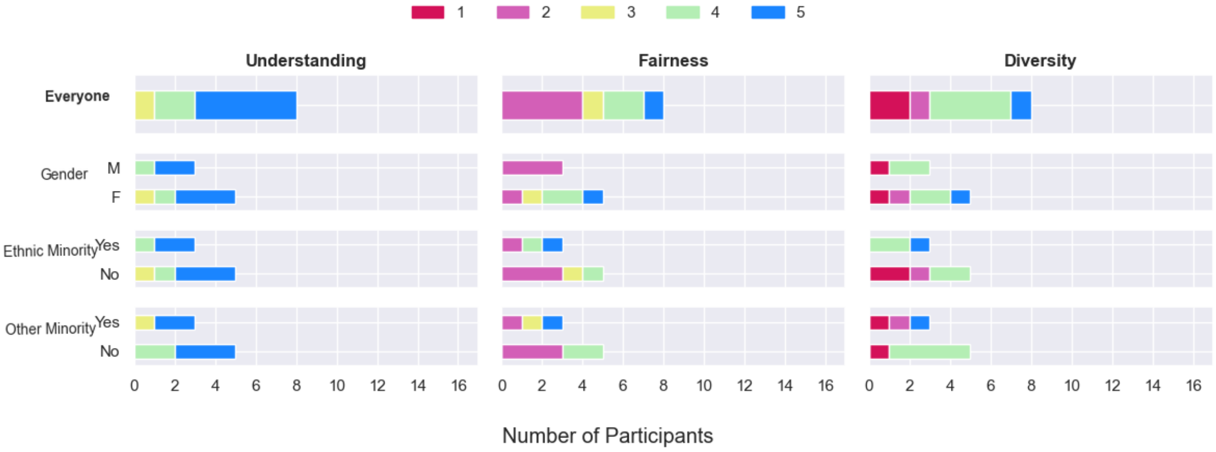


Figure 14: **DI** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of DI, along with the distribution by *gender, ethnic minority* and *other minority* groups.

5.6.1 What is their impression of DI?

Seven out of eight participants rated their understanding of DI at 4 or higher, which is similarly seen across genders and minorities. They understood the notion and were able to express their likes and dislikes about using DI. Two participants, out of the eight, who were interviewed on DI said that it could make organizations accountable by actively trying to improve diversity (E15, E17). *"By having a target that you're actually working towards this early in selection, you're making it more fair for women. A lot of organizations don't want to have these targets in place because you are then held accountable."* - E15. Three other participants expressed their disagreement or dislike for the notion because merit being overlooked for diversity (E11, E16) or minority having lower selection rate than majority (E13). *"I don't like the principle at all. That's saying that if you pay $1 you can pay the woman $0.80 because she's a woman."* - E16. *"If you just have three random females because you sit on gender and not on qualifications. Then it's not a fair system."* - E11.

5.6.2 What is their perception on fairness and improving diversity on using DI?

Half of the participants rated fairness at 2 whereas, half of the participants rated diversity at 4. Two participants called the notion fair when organizations are working towards certain diversity goals (E14, E15). *"This is a measurable target in place to ensure that you are working towards gender fair representation in your organization."* - E15. *"If percentage of women is still less than the percentage of men [in organizations], then it's fair. It's again when it's necessary for the organization and they are equally suitable, but there has to be reason and the reason should be explainable."* - E14. Three other participants considered the notion to be unfair (E10, E11, E13). While, one participant said that it was to a large extent based on a sensitive feature(gender), the other two participants said that it was unfair because minority had lower selection rate compared to the majority. *"It doesn't make sense because you're selecting people on gender. It's not a fair way to manage processes."* - E11. *"It's not fair. It's because like I said, I mean in a tech job or something, women have a harder time getting there anyway. So you should prevent the selection rate for women being lower than for men."* - E13. It is interesting to note from the ratings that all men rated fairness of DI at 2.

Three participants said that the notion could help improve diversity only within some contexts in organizations (E12, E16, E17), while one participant said that despite the notion being unfair, it had the potential of improving diversity (E11). *"If you just have these benchmarks in your in your selection procedure, then yes this will help."* - E12. *"It's not the way we would use. It's not fair, but it would contribute to some improvement in diversity"* - E11.

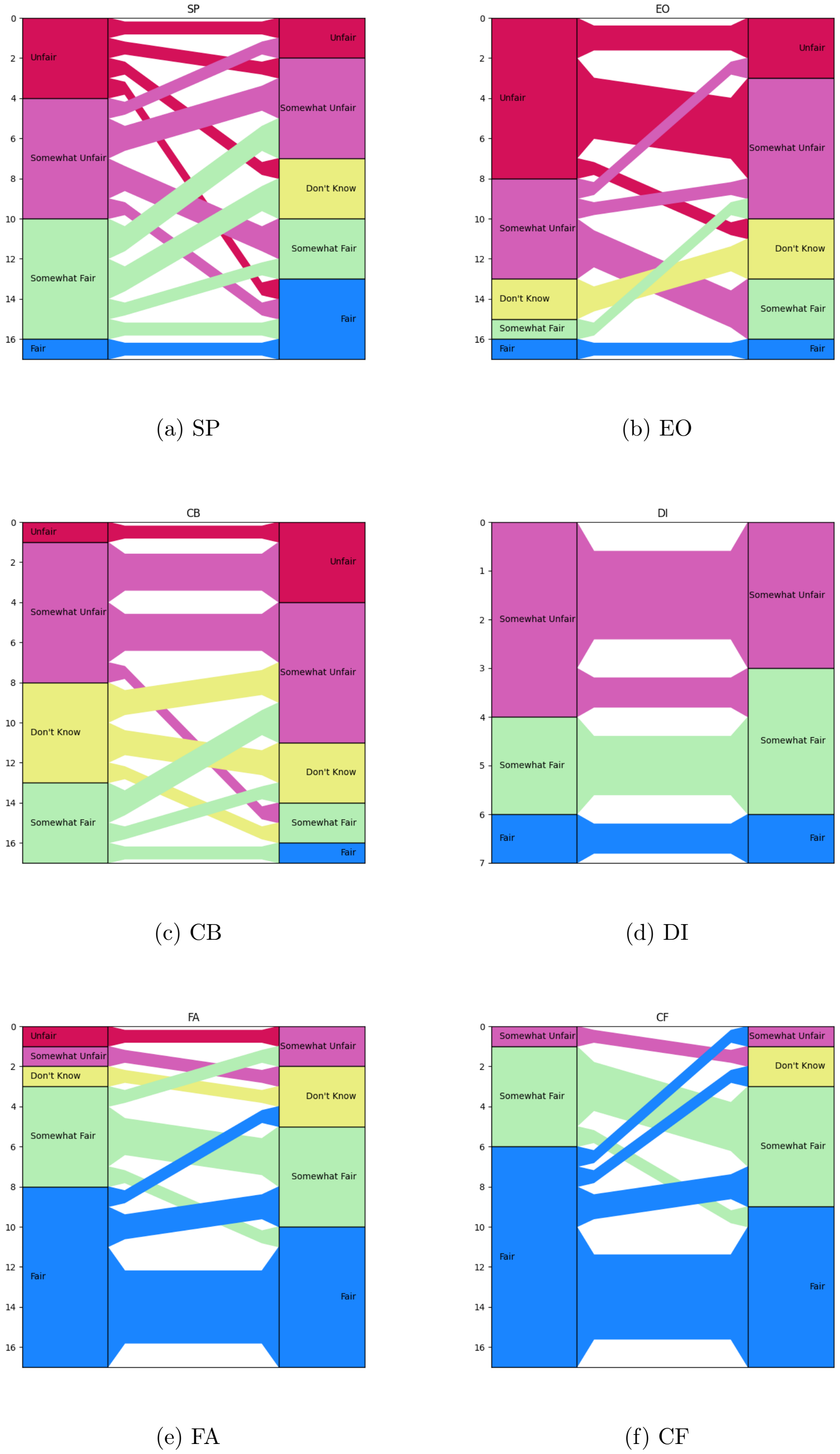
5.6.3 What is the applicability of DI?

Many participants felt that the notion is quite mathematical and seemed concerned about not being able to select the best applicants (E11, E12, E17). *"It's not about filling the numbers. I mean, you could argue statistically it helps you because you have more chance of hiring a female but decisions don't get made on that. It's not a statistical matter. It's a qualitative matter. Your approach is very quantitative. That's a problem"* - E11. *"It's still based on statistics. You're not selecting the best candidates with this."* - E12. *"I assume that when you say, at least 80% selection rate for women, that you select women with a good resume and not just random women."* - E17. One participant said that the notion was not applicable because higher number of selected women does not necessarily guarantee their hiring. *"That means maybe working with recruiters and other people to get those people in the system, because if you don't do that, you'll want to hire a male, even though maybe there's ten females and one male. It's not going to increase the chances of the female to get hired if the male is better than the females."* - E11. On the contrary, another participant said that this notion is applicable in sectors dominated by majority, by increasing the chances of the minority. *"It could help the specifically if you have a very male dominated job and you don't really have that many female applicants. This way of thinking could help."* - E17.

5.7 Changes in the ratings

Having discussed and rated all the notions, participants were asked to rate the same notions on Fairness (F) again. This second round of ratings was obtained to mitigate order effects, where participants did not have access to view their initial ratings. The Figure 15 shows the initial rating and the final ratings given by the participants for each fairness notion.

- **SP:** While we see that majority of the participants changed their rating to either a higher or a lower value, the number of participants calling the notion fair stays the same. This could be due to the fact that half of the participants started the discussion with SP and having reflected on the notions, adjusted their ratings. The final ratings, see proportionally equal number of responses throughout the scale.
- **EO:** The ratings show that participants felt EO was more fair compared to their initial reflection. However, the final ratings still show that majority of the participants gave a rating of 1 or 2.
- **CB:** The overall change shows the majority of the participants gave a rating of 1 or 2, showing that difficulty to grasp the notion affected their rating on fairness.
- **DI:** Out of the eight participants who were interviewed for DI, only two participants changed their ratings from 2 to 4. From the smaller sample of participants, we see that the ratings remained similar in the second round.
- **FA:** The overall effect is decrease in the number of people who rated 5 and increase in the number of people who rated 2 or 3. Having agree with the notion in the first round, participants seemed to adjust their ratings with respect to all the notions.
- **CF:** The changes were similar to FA, where there was a decrease in the number of people who rated 5.



(a) SP

(b) EO

(c) CB

(d) DI

(e) FA

(f) CF

Figure 15: Alluvium plot showing changes in the ratings for fairness for 6 fairness notions. Initial ratings on the *left* are moved to the final ratings on the *right*

6 DISCUSSION

In this section, we present our discussion by connecting the ratings to the reflections of the participants. Moreover, we present additional findings, which were not part of the research questions but provide valuable insights to evaluate the current research and direct towards future research.

6.1 Connecting the ratings and reasons for all the fairness notions

On analyzing the results of the three ratings given for understanding, fairness and diversity, and their corresponding reflection for all the notions, we find that connecting the ratings to the reasons is required to understand the bigger picture and answer the research question. This section connects the participants' ratings for understanding, fairness and diversity to their responses for impression, perception of fairness, perception of diversity and applicability.

6.1.1 Some notions are difficult to understand than others

Amongst the group fairness notions, participants were able to understand the concept of SP, EO and DI. However, CB was difficult to grasp. This was also seen during the design of the toy examples and illustrations, where lab members had deemed it 'too difficult' to understand. Consequently, this lower rating for understandability reflected uncertainty or lower ratings for fairness and diversity. Moreover, this result relates to the study in [121], where participants had difficulty judging complicated fairness notions and ended up selecting SP as the fair option in comparison to a complicated notion. In our study, participants gave ratings for each notion independently and being experts in their domain, were able to reflect on various aspects such as merit and organizational goals. Lastly, participants were also able to understand the concept of FA and CF.

6.1.2 It is difficult to rate the fairness of mathematical notions

Most participants found it quite difficult to provide a rating on fairness. *"Can you call a process that focuses on education and experience in itself fair? I think that's a difficult question to answer"* - E5. *"I find it quite difficult answering your questions. I mean, I would like to separate the fairness and the judging"* - E13. *"It's quite difficult to say if it's fair or not because you think the answer is yes, but I don't think it is".* - E16. Participant's took time to reflect on multiple aspects to help contextualize fairness for themselves. *"It depends how you define fairness. Yeah. It's depends what you understand on the word fairness. Okay. So I'd say don't know because I don't know what you mean by fairness, because fairness is a, is a subjective term. So objectively is not fair. Subjectively it's fair. Depends how you want to view it."* - E11. This indicates the need to contextual fairness within a specific scope in future studies due to the subjective nature of fairness. A similar result [54] showed that lay people did not have any consensus on fairness and their concerns went beyond topics of discrimination to educational background, familial status, equity, etc. This means that more work in needed to translate complicated user friendly definitions to stakeholder friendly language.

6.1.3 Fairness and Diversity considerations may not go hand in hand

For individual fairness notions, many participants took time to suggest that while the notion was fair, it would not help diversity goals of their organizations. *"it's the most fair thing to do, but it doesn't make any sense"* - E1. The tension between fairness and diversity is reflected in the high ratings for fairness but low ratings for diversity. *"If you want to improve gender* *balance in an organization, then sometimes you cannot have this fair criterion. Sometimes you have to let that go."* - E9. *"In general, that will be fair. But if you want to change, unbalanced to balanced, then the decision can be changed."* - E7. For the group fairness notions participants found it difficult to separate their ratings for fairness and diversity. *"It's more than just historical [injustice]. It's also about diversity. We know that having more diverse teams, leads to more creativity and different ideas. So we don't want to create a strict view of one type of candidate that meets like a mold that."* - E11. *"If it's not diverse and gender equal, it's not fair."* - E1. This points towards a need to look at holistic view of fairness and diversity because the fairness notions may not be advancing with societal needs regarding diversity. This aligns with literature suggesting looking beyond discrimination [54] because different scopes impact how people perceive fairness [84]. In our study, we uncover that diversity is an important consideration when addressing fairness in hiring.

6.1.4 Fairness notions in their current state are limited in their applicability

Participants expressed numerous concerns about the applicability of each of the six fairness notions. The major concern noted for SP, EO , FA and CF was the inability of the notion to help improve diversity because majority-minority status quo would be be affected by using these notions. While this concern was expressed in relation to equal selection rates for SP and EO, it was in relation to every individual being defined by different characteristics for FA and CF. *"I presume that you want equal representation and you can't do that by focusing on percentages because as you see the end of the funnel, you'll still end up with a majority and a minority."* - E5. *""It could be that a male applicant has more profit of this fairness notion than a woman because maybe a woman was pregnant a couple of times".* - E3. Moreover, participants expressed that a skewed applicant pool with fewer or no minorities, makes it difficult to put notions such as SP, EO and DI into practice. The applicability of fairness notions, though used in practice [111], is under-researched in literature, pointing towards the need for collaboration between different stakeholders when designing fairness promoting directives in organizations.

6.2 Additional Findings

The semi-structured interviews allowed the participants to provide insights not intended by the research questions. However, these additional findings shed more light on deciphering the perception of fairness and diversity of the mathematical notions. The main themes prevalent across interviews are discussed next.

6.2.1 Use of actual outcome raises many doubts and concerns

When shown the actual outcome as decisions made by a previous committee, most of the participants questioned it's use and validity in the current decisions. *"every hiring committee thinks of themselves as being fair. But it turns out that often they're not, because it's so hard to judge things like."* - E13. *"They are to me unfair because they are so highly at the risk of being influenced by previous people that you don't know what was their rationale behind a certain decision."* - E10. Specifically, for EO participants said that they need space to disagree with the actual outcome because it could perpetuate bias. *"If you ask me what happened and and why is this like this, why did they do this then? That part I don't understand."* - E17. *""Instead of A1, I think I would like to interview A7 or A10 only because I'm very curious to see what they're about and just to find out if there is anything that the previous selection decisions [missed]."* - E16. However, when reflecting on CB, the usage of the actual outcome brought concerns about transparency. *"" If I don't understand it, I can't see how it's going to help me."* - E6. As actual outcome comes from the data driving algorithmic systems, this highlights concerns about data quality. With ramifications on diversity goals of organizations, transparency is needed on what data is collected and what is measures because it can affect the users' trust in the system [117].

6.2.2 There is discomfort in weighing merit and diversity in selections

Many participants said that merit was important and diversity considerations should not precede merit. *"It doesn't serve justice to what you want to achieve in the end, right? What you want is the best candidate and hopefully it should be also a female."* - E11. *"General rule is you try to select as much as you can on merit and you only do this [fairness notions] if you really want to choose for gender balance".* - E7. Some important tensions seen throughout the discussion with participants was whether to include a sensitive feature such as gender so explicitly in the selection process. *"I don't think it's fair to select on gender."* - E17. *"I think that's not really fair. When you when it's, for instance, the first thing you look at when you are starting the hiring procedure."* - E4. Another concern that emerged was the fear of selecting unqualified female candidates for the 'sake' of fairness. *"It feels a lot like the woman is there as an excuse."* - E10. *"You should always select based on merit and merit doesn't come presented in gender."* - E1.

However, examining this theme according to the gender of the participants, highlights an interesting point. While most participants deliberated both the concerns on gender and merit, many women explicitly expressed that general notion of merit favored men and more inclusive takes on merits are needed before such notions could be applicable and called fair. *"We see also that women who, if they have children in their careers, that it has an impact on maybe a number of publications or whatever, and that should not hinder them in making progress or being hired because male applicant has maybe more publications".* - E8. *"I find this topic is just very complex because when you look at experience, we see that because systematically, women get less chances in work than men do, the women also have less experience and therefore, have less strong resumes. But that is a system and you want to break through that system because we always say fix the system, not the women".* - E2.

6.2.3 Domain-specific context and critical nuances are missing in mathematical fairness notions

Many participants expressed that fairness notions miss many critical nuances and context to questions of early selection. Not having access to these nuances and context, such as team composition, type of job role, type of industry, size of the organization and it's goals, hinders them from being certain of their answers, and makes them more resistant to see fairness notions as potential ways to get more explicit and standardized policy. *"The most important thing is always context. Because you can't say that something is fair without having something to bounce it off of. Like the company or the status quo of the company, the hiring process itself. There are many more factors than just equality, than just minority majority or male female. It should be more dynamic and not like a fixed way of approaching [fairness]"* - E6. *"It needs more clarity. So if you say within the context of the needs of [diversity], it's fair. If you would say in context of the male candidate, it would probably be unfair."* - E11. More effort will need to be spent on connecting these notions further with the domain. Here, it is likely that a 'one size fits all' approach will not exist. *"The definition of fairness depends on the circumstances, the environment you're in. Because if it's been an unequal system, what is fair might be different from when it is an equal system."* - E5.

6.2.4 Minorities may have different reflections on fairness

Some participants from a minority background refused to answer the questions, saying that they were forced to rate an option they did not agree with. “So it’s it’s [sic] a bit of a dilemma. There’s no there’s no clear answer on this. So you’re trying to force me into a box and you can’t." - E11. “I’m missing quite a bit of context. It’s really hard. I can’t. I can’t give a proper answer. I find it really hard. These kind of notions, the way we are thinking about this diversity and inclusion. You need to have these numbers, right? Because then you have a better understanding. But for me it’s too simple, too much simplified actually" - E6. In our study, most of the professionals we reached did not identify as belonging to a minority. In the future, it will be interesting to recruit a balanced sample appropriately representing different minority groups to investigate potential response differences. To further strengthen a case for this, for the sensitive attribute on which we had a reasonably balanced sample (gender), on closer examination, we see that proportionally more women compared to men were skeptical about all the notions’ abilities to improve diversity. This direction is aligned with the work in [129], [22], which shows how algorithmic fairness evokes negative emotions from minority groups regarding racial and economic justice and the importance to

6.2.5 Ratings on fairness, do not necessarily reflect sentiment

Even if a larger sample of participants and ratings would be reached, it is important to not only look at numerical ratings, but also at the rationale behind them. Most of the participants felt that they could not disagree with individual fairness notions of FA and CF and rated it as fair. However, their language suggested some restrain or skepticism. “The principal is fair, but the human brain is not fair.So yes, I’m sorry I find is a very hard question. I suppose the principle is fair." - E10. Sentiments of disinterest in continuing the discussion with all the notions were observed during the interviews, despite participants being enthusiastic on talking about such topics, reflects that the ratings do not convey the full picture. “if you start doing this, is it fair? But the notion and mindset of fairness needs to change. You need to first understand who is the minority, do they actually get the same opportunity? No." - E8. “in the perfect world where everybody would be treated equally and we have equal opportunities, then this principle I think is great but we are not in that world right now." - E10.

6.2.6 Fairness notions are not a replacement towards efforts towards fairness in hiring

Many participants liked the concept behind the fairness notions, but said that more work in different stages of the hiring pipeline is needed for fair and diverse hiring. *“I’m really a big fan of making these notions so that you help your organization or hiring managers to decrease their bias. So they have a lot of potential. I like these notions, but they need some improvements."* - E10. Many participants expressed the need to improve the stages involving interviews or interaction with job applicants, saying human biases need to be overcome. *“So in at our [organization], it is obligatory by now to do a gender diversity workshop if you want to be in a hiring committee which I think is very good."* - E13. *“[we] do a stigma awareness training. It’s just telling you that there are things that happen in society that do not help the inclusion talk"*. - E16. Whereas, some participants suggested working in the first stages by attracting the suitable candidates. *“that means maybe working with recruiters and other people to get those people in the system"* - E11. “if you start doing this, is it fair? But the notion and mindset of fairness needs to change. You need to first understand who is the minority, do they actually get the same opportunity? No. It should be about equity." - E8. This shows that mathematical interventions, in their current state are insufficient and needs collaboration between multiple stakeholders to define policies on fairness in hiring.

With this comprehensive discussion, we move towards answering the main research question in the next section.

7 CONCLUSION

We conclude the thesis report by answering the main research question, presenting the limitations of the study, directions for future work, describing the final conclusion of the research.

7.1 Answer to the research question

With the analysis concrete with the results and discussion, we answer the main research question.

RQ : How do organizational representatives understand and perceive different mathematical fairness notions in the context of early candidate selection in hiring?

Answer : While organizational representatives conceptually understand most mathematical fairness notions, they find it difficult to judge them regarding fairness and diversity due to a lack of domain-specific context declaring that they miss critical nuances that hinder applicability in the current world. The quantitative notions do not necessarily allow reaching their organizational goals of equity, diversity, and inclusion, making them resistant to seeing mathematical fairness notions as potential approaches toward standardized policy.

7.2 Limitations and Future work

We identified several limitations in our study. First, the research participants were limited to professionals in the Netherlands, which is not representative of the country's demographics due to the nature of the study. The interviewed professionals also work with skilled job applicants. This indicates that the results are not necessarily transferable across different cultures, demographics, or varying skill sets, which may require more study.

Second, our design set-up handled a single binary valued sensitive feature, a small fictional dataset and a synthetic machine learning model. These choices were sufficient for the study. The choice of a single binary valued variable, *gender* was adapted from the initial domain interviews as the only feasible option for organizations to track. A small fictional dataset was used to ease the discussion with organizational representatives, enabling them to focus on the topic of fairness. Due to the small fictional dataset, training a real algorithm was impossible. All of the above suggests that the applicability of our study can differ when multiple or intersectional sensitive values are used from a large dataset to train a machine learning algorithm, towards which we direct future work.

Third, during our research, we also interviewed a Bachelor and a Master admission coordinator, whose reflections suggested similarities with the hiring domain. However, upon closer examination, we discovered many differences such as using a standardized test for grading performance, organizational goal of study success, and narrowing from a large to a small subset of students for admissions, which deems it a different topic of study.

Lastly, we presented the participants with only our iterative finalized toy examples. Participants being confronted with questions of fairness based on mathematical concepts for the first time did not leave enough room to discuss different designs. For future work, it would be interesting to study the effect of different examples and explanation styles on the perception of fairness of mathematical notions.

7.3 Conclusion

In this study, we interviewed 17 professionals from executive functions, talent acquisition, HR, I/O psychology, and diversity and inclusion operations in The Netherlands. By designing user-friendly illustrations and explanations in the context of early candidate selection in hiring, we explore their ratings and responses to four group and two individual fairness notions. Our qualitative investigation suggests that these fairness notions raise three concerns. One, they lack additional contexts such as a company's size or diversity goals. Two, they give rise to several ethical and practical concerns such as lack of trust in the data, disadvantaging minorities, or the selection of unqualified applicants. Lastly, they act only as a small step towards fairness in the large hiring pipeline. We conclude that a qualitative approach in collaboration between designers, practitioners, and policymakers is the key to refinement and contextualization of future technologically enabled fair hiring policies. Our participants' intrinsic motivation to engage with the topic of fairness strengthens our case.

References

[1] Anita S Acharya et al. “Sampling: Why and how of it”. In: *Indian Journal of Medical Specialties* 4.2 (2013), pp. 330–333.

[2] Alekh Agarwal et al. “A Reductions Approach to Fair Classification”. In: *arXiv:1803.02453 [cs]* (July 16, 2018). arXiv: 1803.02453. URL: http://arxiv.org/abs/1803.02453 (visited on 01/24/2022).

[3] Ifeoma Ajunwa. “An Auditing Imperative for Automated Hiring”. In: (2019).

[4] Julia Angwin et al. “Machine bias”. In: *Ethics of Data and Analytics*. Auerbach Publications, 2016, pp. 254–264.

[5] Jeremy Ashkenas, Haeyoun Park, and Adam Pearce. “Even with affirmative action, Blacks and Hispanics are more underrepresented at top colleges than 35 years ago”. In: *The New York Times* 2 (2017).

[6] Carolyn Ashurst et al. “Why Fair Labels Can Yield Unfair Predictions: Graphical Conditions for Introduced Unfairness”. In: (), p. 10.

[7] Arthur Asuncion and David Newman. *UCI machine learning repository*. 2007.

[8] Stijn Baert. “Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005”. In: (2005), p. 26.

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. “Fairness and Machine Learning”. In: (), p. 253.

[10] Solon Barocas and Andrew D Selbst. “Big data’s disparate impact”. In: *Calif. L. Rev.* 104 (2016), p. 671.

[11] Richard Berk et al. “A Convex Framework for Fair Regression”. In: *arXiv:1706.02409 [cs, stat]* (June 7, 2017). arXiv: 1706.02409. URL: http://arxiv.org/abs/1706.02409 (visited on 01/21/2022).

[12] Richard Berk et al. “Fairness in Criminal Justice Risk Assessments: The State of the Art”. In: *arXiv:1703.09207 [stat]* (May 27, 2017). arXiv: 1703.09207. URL: http://arxiv.org/abs/1703.09207 (visited on 03/07/2022).

[13] Marianne Bertrand and Sendhil Mullainathan. “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination”. In: *THE AMERICAN ECONOMIC REVIEW* (2004), p. 24.

[14] Peter J Bickel, Eugene A Hammel, and J William O’Connell. “Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.” In: *Science* 187.4175 (1975), pp. 398–404.

[15] Dan Biddle. *Adverse impact and test validation: A practitioner’s guide to valid and defensible employment testing*. Routledge, 2017.

[16] Reuben Binns. “Fairness in Machine Learning: Lessons from Political Philosophy”. In: (), p. 11.

[17] Reuben Binns. “On the apparent conflict between individual and group fairness”. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 514–524.

[18] Reuben Binns et al. *’It’s Reducing a Human Being to a Percentage’; Perceptions of Justice in Algorithmic Decisions*. preprint. SocArXiv, Jan. 31, 2018. DOI: 10.31235/osf.io/9wqxr. URL: https://osf.io/9wqxr (visited on 06/08/2022).

[19] J Stewart Black and Patrick van Esch. "AI-enabled recruiting: What is it and how should a manager use it?" In: *Business Horizons* 63.2 (2020), pp. 215–226.

[20] Miranda Bogen and Aaron Rieke. "Help wanted: An examination of hiring algorithms, equity, and bias". In: (2018).

[21] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3.2 (Jan. 2006), pp. 77–101. ISSN: 1478-0887, 1478-0895. DOI: 10.1191/1478088706qp063oa. URL: http://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa (visited on 07/21/2022).

[22] Anna Brown et al. "Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19: CHI Conference on Human Factors in Computing Systems. Glasgow Scotland Uk: ACM, May 2, 2019, pp. 1–12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300271. URL: https://dl.acm.org/doi/10.1145/3290605.3300271 (visited on 06/14/2022).

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

[24] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. "Balanced neighborhoods for multi-sided fairness in recommendation". In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 202–214.

[25] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. "Building classifiers with independency constraints". In: *2009 IEEE International Conference on Data Mining Workshops*. IEEE. 2009, pp. 13–18.

[26] Toon Calders and Sicco Verwer. "Three naive Bayes approaches for discrimination-free classification". In: *Data Mining and Knowledge Discovery* 21.2 (Sept. 2010), pp. 277–292. ISSN: 1384-5810, 1573-756X. DOI: 10.1007/s10618-010-0190-x. URL: http://link.springer.com/10.1007/s10618-010-0190-x (visited on 12/08/2021).

[27] Aaron Chalfin et al. "Productivity and selection of human capital with machine learning". In: *American Economic Review* 106.5 (2016), pp. 124–27.

[28] Tomas Chamorro-Premuzic and Reece Akhtar. *Should Companies Use AI to Assess Job Candidates?* 2017. URL: https://hbr.org/2019/05/should-companies-use-ai-to-assess-job-candidates.

[29] Le Chen et al. "Investigating the impact of gender on rank in resume search engines". In: *Proceedings of the 2018 chi conference on human factors in computing systems*. 2018, pp. 1–14.

[30] Silvia Chiappa. "Path-specific counterfactual fairness". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 7801–7808.

[31] Silvia Chiappa and William S. Isaac. "A Causal Bayesian Networks Viewpoint on Fairness". In: *arXiv:1907.06430 [cs, stat]* 547 (2019), pp. 3–20. DOI: 10.1007/978-3-030-16744-8_1. arXiv: 1907.06430. URL: http://arxiv.org/abs/1907.06430 (visited on 05/05/2022).

[32] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *arXiv:1610.07524 [cs, stat]* (Oct. 24, 2016). arXiv: 1610.07524. URL: http://arxiv.org/abs/1610.07524 (visited on 01/24/2022).

[33] Alexandra Chouldechova and Aaron Roth. “The Frontiers of Fairness in Machine Learning”. In: *arXiv:1810.08810 [cs, stat]* (Oct. 20, 2018). arXiv: 1810.08810. URL: http://arxiv.org/abs/1810.08810 (visited on 01/06/2022).

[34] Kevin A. Clarke. “The Phantom Menace: Omitted Variable Bias in Econometric Research”. In: *Conflict Management and Peace Science* 22.4 (2005), pp. 341–352. URL: https://EconPapers.repec.org/RePEc:sae:compsc:v:22:y:2005:i:4:p:341-352.

[35] Sam Corbett-Davies et al. “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.

[36] Bo Cowgill. “Bias and productivity in humans and machines”. In: *Columbia Business School Research Paper Forthcoming* (2019).

[37] Elliot Creager et al. “Flexibly fair representation learning by disentanglement”. In: *International conference on machine learning*. PMLR. 2019, pp. 1436–1445.

[38] Brian d’Alessandro, Cathy O’Neil, and Tom LaGatta. “Conscientious Classification: A Data Scientist’s Guide to Discrimination-Aware Classification”. In: *Big Data* 5.2 (June 2017), pp. 120–134. ISSN: 2167-6461, 2167-647X. DOI: 10.1089/big.2016.0048. arXiv: 1907.09013. URL: http://arxiv.org/abs/1907.09013 (visited on 03/29/2022).

[39] David Danks and Alex John London. “Algorithmic Bias in Autonomous Systems.” In: *IJCAI*. Vol. 17. 2017, pp. 4691–4697.

[40] William Dieterich, Christina Mendoza, and Tim Brennan. “COMPAS risk scales: Demonstrating accuracy equity and predictive parity”. In: *Northpointe Inc* 7.4 (2016).

[41] Jonathan Dodge et al. “Explaining models: an empirical study of how explanations impact fairness judgment”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19: 24th International Conference on Intelligent User Interfaces. Marina del Ray California: ACM, Mar. 17, 2019, pp. 275–285. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302310. URL: https://dl.acm.org/doi/10.1145/3301275.3302310 (visited on 06/14/2022).

[42] *doi:10.1016/j.rssm.2006.06.001 | Elsevier Enhanced Reader*. DOI: 10.1016/j.rssm.2006.06.001. URL: https://reader.elsevier.com/reader/sd/pii/S0276562406000278?token=6BC7FB4F2BB37DC5329AED5361F0DA2AFE7698696210C33128042E1C60
originRegion=eu-west-1&originCreation=20220221104027 (visited on 02/21/2022).

[43] Cynthia Dwork et al. “Fairness Through Awareness”. In: *arXiv:1104.3913 [cs]* (Nov. 28, 2011). arXiv: 1104.3913. URL: http://arxiv.org/abs/1104.3913 (visited on 01/24/2022).

[44] Shady Elbassuoni et al. *Exploring Fairness of Ranking in Online Job Marketplaces*. Version Number: 1 Type: dataset. 2019. DOI: 10.5441/002/EDBT.2019.77. URL: https://openproceedings.org/2019/conf/edbt/EDBT19_paper_230.pdf (visited on 11/03/2021).

[45] Michael Feldman et al. “Certifying and Removing Disparate Impact”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’15: The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney NSW Australia: ACM, Aug. 10, 2015, pp. 259–268. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2783311. URL: https://dl.acm.org/doi/10.1145/2783258.2783311 (visited on 12/07/2021).

[46] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. “A confidence-based approach for balancing fairness and accuracy”. In: *Proceedings of the 2016 SIAM international conference on data mining*. SIAM. 2016, pp. 144–152.

[47] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks". In: *Fed. Probation* 80 (2016), p. 38.

[48] Paul Frijters. "Discrimination and job-uncertainty". In: *Journal of economic behavior & organization* 36.4 (1998), pp. 433–446.

[49] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. "Fairness testing: testing software for discrimination". In: *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*. 2017, pp. 498–510.

[50] Michael Geden and Joshua Andrews. "Fair and Interpretable Algorithmic Hiring using Evolutionary Many-Objective Optimization". In: (), p. 9.

[51] IJ Goodfellow et al. *Neural Information Processing Systems (NIPS)*. 2014.

[52] Leo A Goodman. "Snowball sampling". In: *The annals of mathematical statistics* (1961), pp. 148–170.

[53] Nina Grgic-Hlacˇa et al. "The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making". In: (), p. 11.

[54] Nina Grgic-Hlaca et al. "Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction". In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. the 2018 World Wide Web Conference. Lyon, France: ACM Press, 2018, pp. 903–912. ISBN: 978-1-4503-5639-8. DOI: 10.1145/3178876.3186138. URL: http://dl.acm.org/citation.cfm?doid=3178876.3186138 (visited on 06/30/2022).

[55] Priyanko Guchait et al. "Video interviewing: A potential selection tool for hospitality managers–A study to understand applicant perspective". In: *International Journal of Hospitality Management* 36 (2014), pp. 90–100.

[56] Philipp Hacker. "Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law". In: *Common Market Law Review* 55.4 (2018).

[57] Sara Hajian and Josep Domingo-Ferrer. "A methodology for direct and indirect discrimination prevention in data mining". In: *IEEE transactions on knowledge and data engineering* 25.7 (2012), pp. 1445–1459.

[58] Sara Hajian and Josep Domingo-Ferrer. "Direct and indirect discrimination prevention methods". In: *Discrimination and privacy in the information society*. Springer, 2013, pp. 241–254.

[59] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *arXiv:1610.02413 [cs]* (Oct. 7, 2016). arXiv: 1610.02413. URL: http://arxiv.org/abs/1610.02413 (visited on 12/08/2021).

[60] Kimberly A Houser. "Can AI solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making". In: *Stan. Tech. L. Rev.* 22 (2019), p. 290.

[61] Han-Yin Huang and Cynthia C. S. Liem. *Social Inclusion in Curated Contexts: Insights from Museum Practices*. May 10, 2022. DOI: 10.1145/3531146.3533095. arXiv: 2205.05192[cs]. URL: http://arxiv.org/abs/2205.05192 (visited on 05/19/2022).

[62] Abigail Z. Jacobs and Hanna Wallach. “Measurement and Fairness”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* FAccT ’21: 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event Canada: ACM, Mar. 3, 2021, pp. 375–385. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445901. URL: https://dl.acm.org/doi/10.1145/3442188.3445901 (visited on 03/10/2022).

[63] Matthew Joseph et al. “Fairness in Learning: Classic and Contextual Bandits”. In: *arXiv:1605.07139 [cs, stat]* (Nov. 7, 2016). arXiv: 1605.07139. URL: http://arxiv.org/abs/1605.07139 (visited on 03/07/2022).

[64] Faisal Kamiran and Toon Calders. “Classification with no discrimination by preferential sampling”. In: *Proc. 19th Machine Learning Conf. Belgium and The Netherlands.* Citeseer. 2010, pp. 1–6.

[65] Faisal Kamiran and Toon Calders. “Classifying without discriminating”. In: *2009 2nd International Conference on Computer, Control and Communication.* 2009 2nd International Conference on Computer, Control and Communication (IC$). Karachi, Pakistan: IEEE, Feb. 2009, pp. 1–6. ISBN: 978-1-4244-3313-1. DOI: 10.1109/IC4.2009.4909197. URL: http://ieeexplore.ieee.org/document/4909197/ (visited on 12/07/2021).

[66] Faisal Kamiran and Toon Calders. “Data preprocessing techniques for classification without discrimination”. In: *Knowledge and Information Systems* 33.1 (Oct. 2012), pp. 1–33. ISSN: 0219-1377, 0219-3116. DOI: 10.1007/s10115-011-0463-8. URL: http://link.springer.com/10.1007/s10115-011-0463-8 (visited on 02/28/2022).

[67] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. “Discrimination Aware Decision Tree Learning”. In: *2010 IEEE International Conference on Data Mining.* 2010 IEEE 10th International Conference on Data Mining (ICDM). Sydney, Australia: IEEE, Dec. 2010, pp. 869–874. ISBN: 978-1-4244-9131-5. DOI: 10.1109/ICDM.2010.50. URL: http://ieeexplore.ieee.org/document/5694053/ (visited on 12/16/2021).

[68] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. “Decision theory for discrimination-aware classification”. In: *2012 IEEE 12th International Conference on Data Mining.* IEEE. 2012, pp. 924–929.

[69] Faysal Kamiran and Indre Zliobaite. “Explainable and Non-explainable Discrimination in Classification”. English. In: *Discrimination and Privacy in the Information Society.* Vol. 3. Studies in Applied Philosophy, Epistemology and Rational Ethics. International: Springer, 2013, pp. 155–170. ISBN: 978-3-642-30486-6. DOI: 10.1007/978-3-642-30487-3_8.

[70] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. “Fairness-aware learning through regularization approach”. In: *2011 IEEE 11th International Conference on Data Mining Workshops.* IEEE. 2011, pp. 643–650.

[71] Maria Kasinidou et al. “I agree with the decision, but they didn’t deserve this: Future Developers’ Perception of Fairness in Algorithmic Decisions”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* FAccT ’21: 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event Canada: ACM, Mar. 3, 2021, pp. 690–700. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445931. URL: https://dl.acm.org/doi/10.1145/3442188.3445931 (visited on 06/16/2022).

[72] Michael Kearns et al. “An Empirical Study of Rich Subgroup Fairness for Machine Learning”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency.* FAT* ’19: Conference on Fairness, Accountability, and Transparency. Atlanta GA USA: ACM, Jan. 29, 2019, pp. 100–109. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287592. URL: https://dl.acm.org/doi/10.1145/3287560.3287592 (visited on 01/18/2022).

[73] Michael Kearns et al. “Preventing Fairness Gerrymandering:Auditing and Learning for Subgroup Fairness”. In: (), p. 9.

[74] Niki Kilbertus et al. “Avoiding discrimination through causal reasoning”. In: *Advances in neural information processing systems* 30 (2017).

[75] Diederik P Kingma and Max Welling. “Stochastic gradient VB and the variational auto-encoder”. In: *Second International Conference on Learning Representations, ICLR.* Vol. 19. 2014, p. 121.

[76] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. “Inherent Trade-Offs in the Fair Determination of Risk Scores”. In: *arXiv:1609.05807 [cs, stat]* (Nov. 17, 2016). arXiv: 1609.05807. URL: http://arxiv.org/abs/1609.05807 (visited on 01/18/2022).

[77] Jon Kleinberg et al. “Human decisions and machine predictions”. In: *The quarterly journal of economics* 133.1 (2018), pp. 237–293.

[78] Alina Köchling and Marius Claus Wehner. “Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development”. In: *Business Research* 13.3 (Nov. 2020), pp. 795–848. ISSN: 2198-3402, 2198-2627. DOI: 10.1007/s40685-020-00134-w. URL: https://link.springer.com/10.1007/s40685-020-00134-w (visited on 04/27/2022).

[79] Kozodoi. *Fairness in credit scoring: Assessment, implementation and profit implications | Elsevier Enhanced Reader.* DOI: 10.1016/j.ejor.2021.06.023. URL: https://reader.elsevier.com/reader/sd/pii/S0377221721005385?token=6AF085CF49C739110893571A87E9A53BCC52A274B7CFA91D34AE169229BB0C0DD3FA9B58A8CE251B06 originRegion=eu-west-1&originCreation=20220418093244 (visited on 04/18/2022).

[80] Matt J Kusner et al. “Counterfactual fairness”. In: *Advances in neural information processing systems* 30 (2017).

[81] Christina N Lacerenza et al. “Leadership training design, delivery, and implementation: A meta-analysis.” In: *Journal of Applied Psychology* 102.12 (2017), p. 1686.

[82] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. “ifair: Learning individually fair data representations for algorithmic decision making”. In: *2019 ieee 35th international conference on data engineering (icde).* IEEE. 2019, pp. 1334–1345.

[83] *Laws Enforced by EEOC.* https://www.eeoc.gov/statutes/laws-enforced-eeoc. Accessed 27-07-2022.

[84] Min Kyung Lee. “Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management”. In: *Big Data & Society* 5.1 (Jan. 1, 2018). Publisher: SAGE Publications Ltd, p. 2053951718756684. ISSN: 2053-9517. DOI: 10.1177/2053951718756684. URL: https://doi.org/10.1177/2053951718756684 (visited on 06/08/2022).

[85] Min Kyung Lee. “Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management”. In: *Big Data & Society* 5.1 (2018), p. 2053951718756684.

[86] Min Kyung Lee and Su Baykal. "Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17: Computer Supported Cooperative Work and Social Computing. Portland Oregon USA: ACM, Feb. 25, 2017, pp. 1035–1048. ISBN: 978-1-4503-4335-0. DOI: 10.1145/2998181.2998230. URL: https://dl.acm.org/doi/10.1145/2998181.2998230 (visited on 06/08/2022).

[87] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. "A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management that Allocates Donations to Non-Profit Organizations". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17: CHI Conference on Human Factors in Computing Systems. Denver Colorado USA: ACM, May 2, 2017, pp. 3365–3376. ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025884. URL: https://dl.acm.org/doi/10.1145/3025453.3025884 (visited on 06/08/2022).

[88] Bruno Lepri et al. "Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges". In: *Philosophy & Technology* 31.4 (Dec. 2018), pp. 611–627. ISSN: 2210-5433, 2210-5441. DOI: 10.1007/s13347-017-0279-x. URL: http://link.springer.com/10.1007/s13347-017-0279-x (visited on 04/29/2022).

[89] Lan Li et al. "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*. AIES '21: AAAI/ACM Conference on AI, Ethics, and Society. Virtual Event USA: ACM, July 21, 2021, pp. 166–176. ISBN: 978-1-4503-8473-5. DOI: 10.1145/3461702.3462531. URL: https://dl.acm.org/doi/10.1145/3461702.3462531 (visited on 06/14/2022).

[90] Lan Li et al. "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.

[91] Dirk Lindebaum, Mikko Vesa, and Frank Den Hond. "Insights from "the machine stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations". In: *Academy of Management Review* 45.1 (2020), pp. 247–263.

[92] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. "Does mitigating ML's impact disparity require treatment disparity?" In: *arXiv:1711.07076 [cs, stat]* (Jan. 11, 2019). arXiv: 1711.07076. URL: http://arxiv.org/abs/1711.07076 (visited on 01/18/2022).

[93] Francesco Locatello et al. "Challenging common assumptions in the unsupervised learning of disentangled representations". In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

[94] Joshua R. Loftus et al. "Causal Reasoning for Algorithmic Fairness". In: *arXiv:1805.05859 [cs]* (May 15, 2018). arXiv: 1805.05859. URL: http://arxiv.org/abs/1805.05859 (visited on 05/05/2022).

[95] Christos Louizos et al. "The Variational Fair Autoencoder". In: *arXiv:1511.00830 [cs, stat]* (Aug. 9, 2017). arXiv: 1511.00830. URL: http://arxiv.org/abs/1511.00830 (visited on 05/04/2022).

[96] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. "k-NN as an implementation of situation testing for discrimination discovery and prevention". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2011, pp. 502–510.

[97] A Lutz. "Applicants for jobs at the new DC Walmart face worse odds than people trying to get into Harvard". In: *Business Insider* (2013).

[98] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. "On the Applicability of ML Fairness Notions". In: *arXiv:2006.16745 [cs, stat]* (Oct. 19, 2020). arXiv: 2006.16745. URL: http://arxiv.org/abs/2006.16745 (visited on 11/03/2021).

[99] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys* 54.6 (July 2021), pp. 1–35. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3457607. URL: https://dl.acm.org/doi/10.1145/3457607 (visited on 11/10/2021).

[100] Anay Mehrotra, Bary S. R. Pradelski, and Nisheeth K. Vishnoi. "Selection in the Presence of Implicit Bias: The Advantage of Intersectional Constraints". In: *2022 ACM Conference on Fairness, Accountability, and Transparency.* FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM, June 21, 2022, pp. 599–609. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533124. URL: https://dl.acm.org/doi/10.1145/3531146.3533124 (visited on 06/30/2022).

[101] Shira Mitchell et al. "Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions". In: *Annual Review of Statistics and Its Application* 8.1 (Mar. 7, 2021), pp. 141–163. ISSN: 2326-8298, 2326-831X. DOI: 10.1146/annurev-statistics-042720-125902. arXiv: 1811.07867. URL: http://arxiv.org/abs/1811.07867 (visited on 11/17/2021).

[102] Razieh Nabi and Ilya Shpitser. "Fair Inference on Outcomes". In: (), p. 10.

[103] Iftekhar Naim et al. "Automated analysis and prediction of job interview performance". In: *IEEE Transactions on Affective Computing* 9.2 (2016), pp. 191–204.

[104] Alexandra Olteanu et al. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries". In: *Frontiers in Big Data* 2 (2019). ISSN: 2624-909X. DOI: 10.3389/fdata.2019.00013. URL: https://www.frontiersin.org/article/10.3389/fdata.2019.00013.

[105] Osonde A Osoba and William Welser IV. *An intelligence in our image: The risks of bias and errors in artificial intelligence.* Rand Corporation, 2017.

[106] Judea Pearl. *Causality.* Cambridge university press, 2009.

[107] Judea Pearl et al. "Models, reasoning and inference". In: *Cambridge, UK: CambridgeUniversityPress* 19 (2000), p. 2.

[108] John Platt et al. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74.

[109] Geoff Pleiss et al. "On Fairness and Calibration". In: *Advances in Neural Information Processing Systems.* Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html (visited on 05/02/2022).

[110] Lincoln Quillian et al. “Meta-analysis of field experiments shows no change in racial discrimination in hiring over time”. In: *Proceedings of the National Academy of Sciences* 114.41 (Oct. 10, 2017), pp. 10870–10875. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1706255114. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1706255114 (visited on 02/21/2022).

[111] Manish Raghavan et al. “Mitigating bias in algorithmic hiring: evaluating claims and practices”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* FAT* ’20: Conference on Fairness, Accountability, and Transparency. Barcelona Spain: ACM, Jan. 27, 2020, pp. 469–481. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372828. URL: https://dl.acm.org/doi/10.1145/3351095.3372828 (visited on 01/28/2022).

[112] Peter A. Riach and Judith Rich. “An Experimental Investigation of Sexual Discrimination in Hiring in the English Labor Market”. In: *The B.E. Journal of Economic Analysis & Policy* 6.2 (Jan. 13, 2006). Publisher: De Gruyter. ISSN: 1935-1682. DOI: 10.2202/1538-0637.1416. URL: https://www.degruyter.com/document/doi/10.2202/1538-0637.1416/html (visited on 02/21/2022).

[113] Dan-Olof Rooth. “Automatic associations and discrimination in hiring: Real world evidence”. In: *Labour Economics* 17.3 (June 1, 2010), pp. 523–534. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2009.04.005. URL: https://www.sciencedirect.com/science/article/pii/S0927537109000451 (visited on 02/21/2022).

[114] Filippo Santoni de Sio and Jeroen Van den Hoven. “Meaningful human control over autonomous systems: A philosophical account”. In: *Frontiers in Robotics and AI* (2018), p. 15.

[115] Nripsuta Ani Saxena et al. “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.* AIES ’19: AAAI/ACM Conference on AI, Ethics, and Society. Honolulu HI USA: ACM, Jan. 27, 2019, pp. 99–106. ISBN: 978-1-4503-6324-2. DOI: 10.1145/3306618.3314248. URL: https://dl.acm.org/doi/10.1145/3306618.3314248 (visited on 06/06/2022).

[116] T.M. Scanlon. *Moral Dimensions: Permissibility, Meaning, Blame.* Harvard University Press, Jan. 1, 2008. ISBN: 978-0-674-04314-5. DOI: 10.4159/9780674043145. URL: https://www.degruyter.com/document/doi/10.4159/9780674043145/html (visited on 02/19/2022).

[117] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. ““There Is Not Enough Information”: On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency.* FAccT ’22: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM, June 21, 2022, pp. 1616–1628. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533218. URL: https://dl.acm.org/doi/10.1145/3531146.3533218 (visited on 06/30/2022).

[118] Hong Shen et al. “The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency.* FAccT ’22: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM, June 21, 2022, pp. 440–451. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533110. URL: https://dl.acm.org/doi/10.1145/3531146.3533110 (visited on 06/30/2022).

[119] Edward H Simpson. “The interpretation of interaction in contingency tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 13.2 (1951), pp. 238–241.

[120] Peter Spirtes et al. *Causation, prediction, and search.* MIT press, 2000.

[121] Megha Srivastava, Hoda Heidari, and Andreas 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.* KDD ’19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Anchorage AK USA: ACM, July 25, 2019, pp. 2459–2468. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330664. URL: https://dl.acm.org/doi/10.1145/3292500.3330664 (visited on 03/04/2022).

[122] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. “Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring”. In: *arXiv:2012.00423 [cs]* (Apr. 8, 2021). arXiv: 2012.00423. URL: http://arxiv.org/abs/2012.00423 (visited on 12/09/2021).

[123] Harini Suresh and John V Guttag. “A framework for understanding unintended consequences of machine learning”. In: *arXiv preprint arXiv:1901.10002* 2 (2019).

[124] Harini Suresh and John V. Guttag. “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle”. In: *Equity and Access in Algorithms, Mechanisms, and Optimization* (Oct. 5, 2021), pp. 1–9. DOI: 10.1145/3465416.3483305. arXiv: 1901.10002. URL: http://arxiv.org/abs/1901.10002 (visited on 03/22/2022).

[125] Pooya Tabesh. “Who’s making the decisions? How managers can harness artificial intelligence and remain in charge”. In: *Journal of Business Strategy* (2021).

[126] *Tackling discrimination at work.* https://ec.europa.eu/social/main.jsp?catId=158&langId=en. Accessed 27-07-2022.

[127] Margery Austin Turner, Michael Fix, and Raymond J Struyk. *Opportunities denied, opportunities diminished: Racial discrimination in hiring.* The Urban Insitute, 1991.

[128] Tyler J VanderWeele and Miguel A Hernán. “Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs”. In: *American journal of epidemiology* 175.12 (2012), pp. 1303–1310.

[129] Allison Woodruff et al. “A Qualitative Exploration of Perceptions of Algorithmic Fairness”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* CHI ’18: CHI Conference on Human Factors in Computing Systems. Montreal QC Canada: ACM, Apr. 21, 2018, pp. 1–14. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174230. URL: https://dl.acm.org/doi/10.1145/3173574.3174230 (visited on 06/08/2022).

[130] Blake Woodworth et al. “Learning Non-Discriminatory Predictors”. In: *arXiv:1702.06081 [cs]* (Nov. 1, 2017). arXiv: 1702.06081. URL: http://arxiv.org/abs/1702.06081 (visited on 01/19/2022).

[131] Depeng Xu et al. “Fairgan: Fairness-aware generative adversarial networks”. In: *2018 IEEE International Conference on Big Data (Big Data).* IEEE. 2018, pp. 570–575.

[132] Lynette Yarger, Payton Fay Cobb, and Bikalpa Neupane. “Algorithmic equity in the hiring of underrepresented IT job candidates”. In: *Online Information Review* 44.2 (Jan. 1, 2019). Publisher: Emerald Publishing Limited, pp. 383–395. ISSN: 1468-4527. DOI: 10.1108/OIR-10-2018-0334. URL: https://doi.org/10.1108/OIR-10-2018-0334 (visited on 04/29/2022).

[133] Bianca Zadrozny and Charles Elkan. “Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers”. In: *Icml.* Vol. 1. Citeseer. 2001, pp. 609–616.

[134] Muhammad Bilal Zafar et al. “Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment”. In: *Proceedings of the 26th International Conference on World Wide Web* (Apr. 3, 2017), pp. 1171–1180. DOI: 10.1145/3038912.3052660. arXiv: 1610.08452. URL: http://arxiv.org/abs/1610.08452 (visited on 01/24/2022).

[135] Muhammad Bilal Zafar et al. “Fairness Constraints: Mechanisms for Fair Classification”. In: *arXiv:1507.05259 [cs, stat]* (Mar. 23, 2017). arXiv: 1507.05259. URL: http://arxiv.org/abs/1507.05259 (visited on 01/19/2022).

[136] Richard Zemel. “Learning Fair Representations”. In: (), p. 9.

[137] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. “Mitigating unwanted biases with adversarial learning”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 335–340.

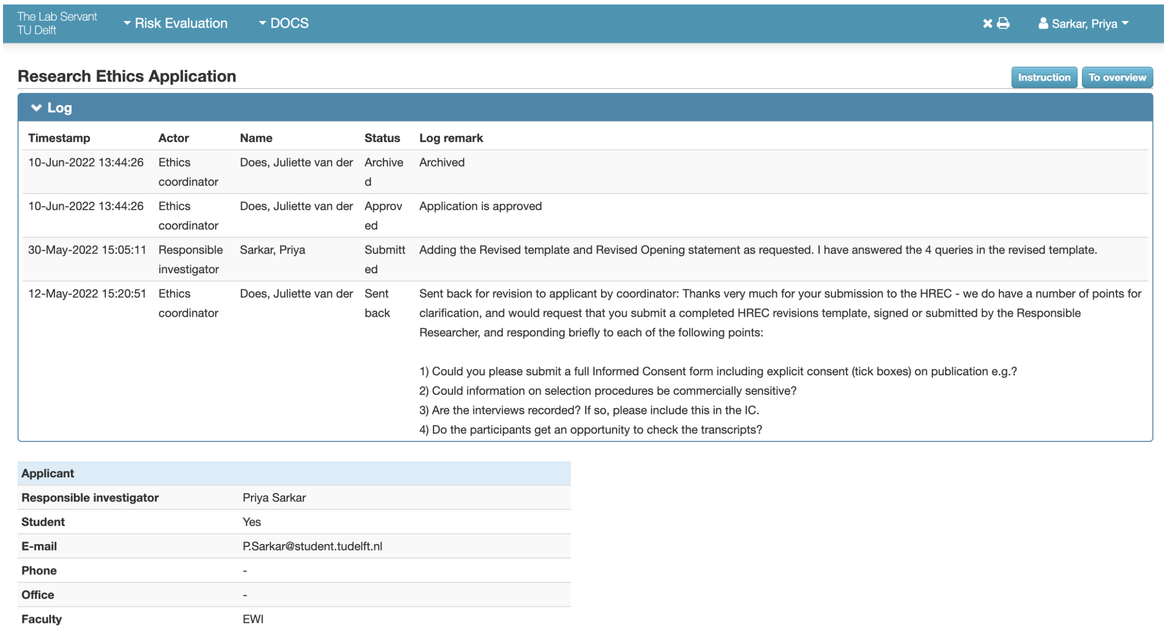
[138] Lu Zhang, Yongkai Wu, and Xintao Wu. “Achieving non-discrimination in data release”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 1335–1344.

[139] Indre Zliobaite. “A survey on measuring indirect discrimination in machine learning”. In: *arXiv:1511.00148 [cs, stat]* (Oct. 31, 2015). arXiv: 1511.00148. URL: http://arxiv.org/abs/1511.00148 (visited on 12/15/2021).

Appendix

A. HREC Approval

The research was approved by HREC as shown below.



Research Ethics Application

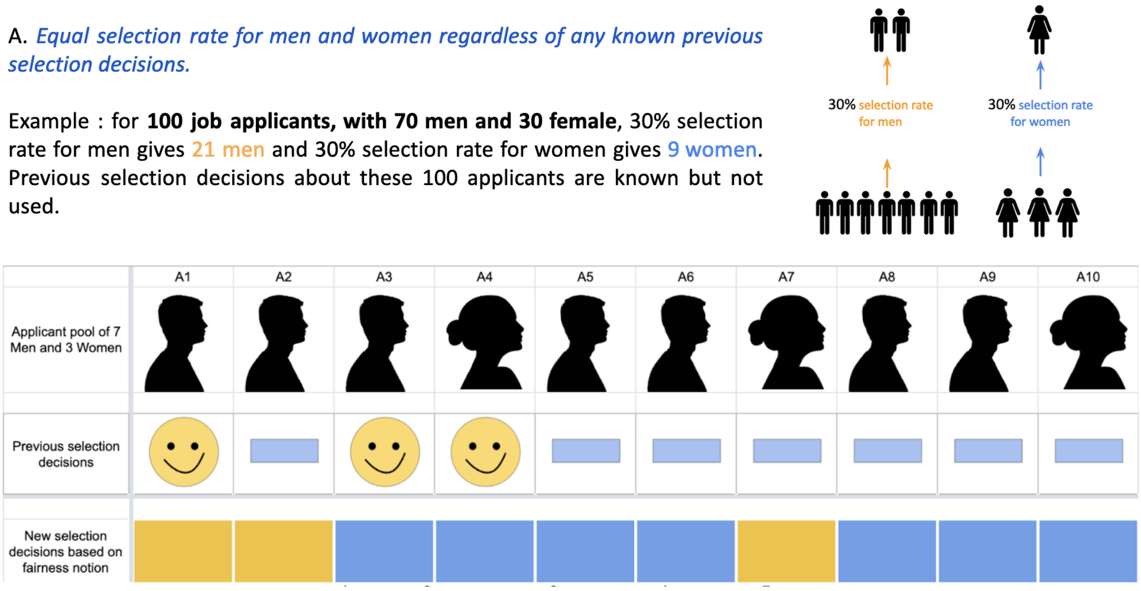
Log

Timestamp	Actor	Name	Status	Log remark
10-Jun-2022 13:44:26	Ethics coordinator	Does, Juliette van der	Archived	Archived
10-Jun-2022 13:44:26	Ethics coordinator	Does, Juliette van der	Approved	Application is approved
30-May-2022 15:05:11	Responsible investigator	Sarkar, Priya	Submitted	Adding the Revised template and Revised Opening statement as requested. I have answered the 4 queries in the revised template.
12-May-2022 15:20:51	Ethics coordinator	Does, Juliette van der	Sent back	Sent back for revision to applicant by coordinator: Thanks very much for your submission to the HREC - we do have a number of points for clarification, and would request that you submit a completed HREC revisions template, signed or submitted by the Responsible Researcher, and responding briefly to each of the following points: 1) Could you please submit a full Informed Consent form including explicit consent (tick boxes) on publication e.g.? 2) Could information on selection procedures be commercially sensitive? 3) Are the interviews recorded? If so, please include this in the IC. 4) Do the participants get an opportunity to check the transcripts?

Applicant	
Responsible investigator	Priya Sarkar
Student	Yes
E-mail	P.Sarkar@student.tudelft.nl
Phone	-
Office	-
Faculty	EWI

Figure 16: HREC Approval

B. Final toy illustrations of the six mathematical fairness notions

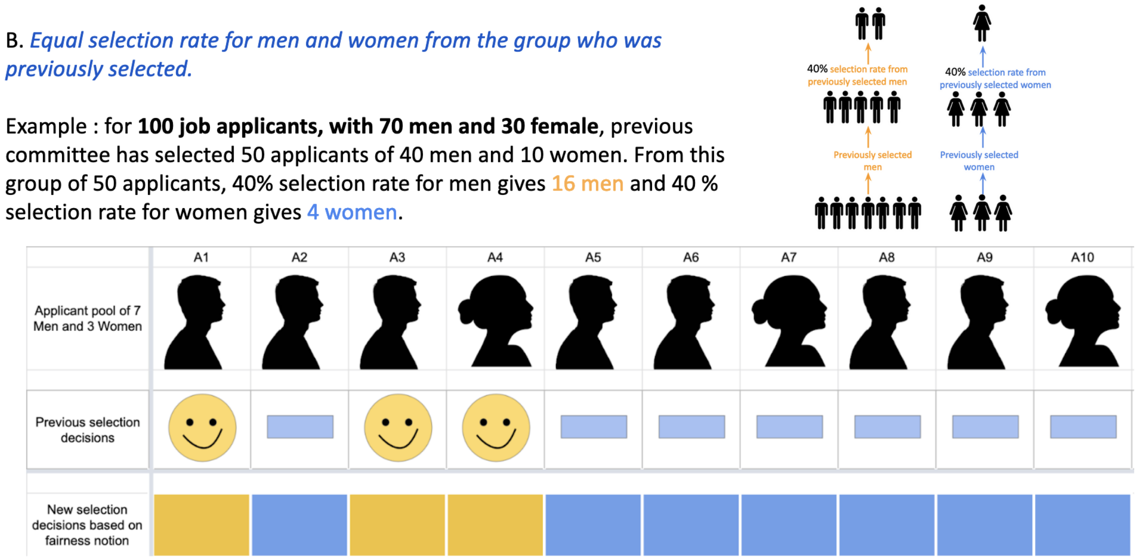


A. *Equal selection rate for men and women regardless of any known previous selection decisions.*

Example : for **100 job applicants, with 70 men and 30 female**, 30% selection rate for men gives 21 men and 30% selection rate for women gives 9 women. Previous selection decisions about these 100 applicants are known but not used.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Applicant pool of 7 Men and 3 Women										
Previous selection decisions	☺	▭	☺	☺	▭	▭	▭	▭	▭	▭
New selection decisions based on fairness notion	🟨	🟨	🟦	🟦	🟦	🟦	🟨	🟦	🟦	🟦

(a) Final design of the toy example for SP



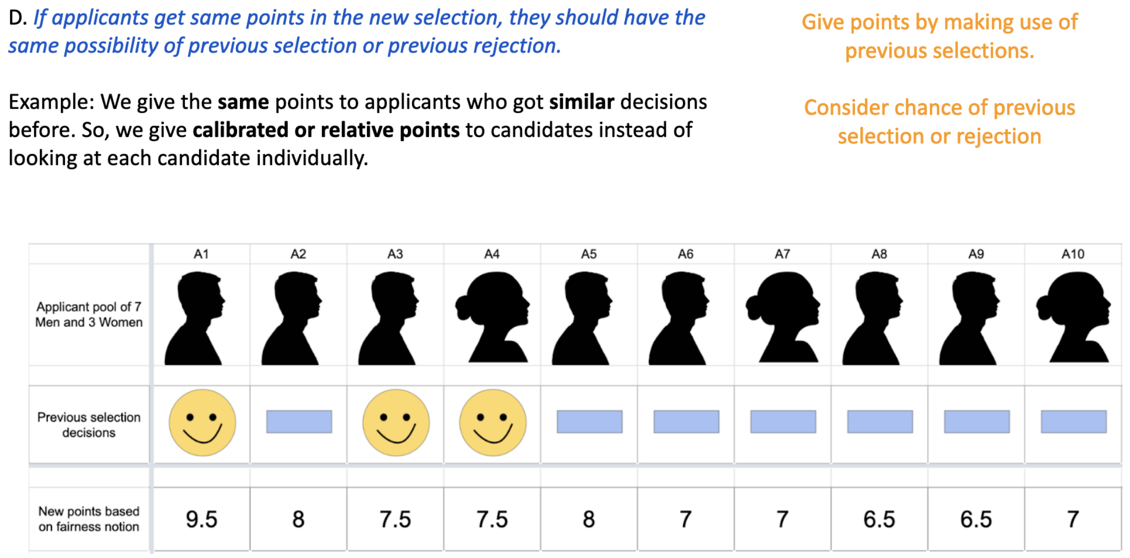
B. *Equal selection rate for men and women from the group who was previously selected.*

Example : for **100 job applicants, with 70 men and 30 female**, previous committee has selected 50 applicants of 40 men and 10 women. From this group of 50 applicants, 40% selection rate for men gives 16 men and 40 % selection rate for women gives 4 women.

40% selection rate from previously selected men
40% selection rate from previously selected women
Previously selected men
Previously selected women

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Applicant pool of 7 Men and 3 Women										
Previous selection decisions	☺	▭	☺	☺	▭	▭	▭	▭	▭	▭
New selection decisions based on fairness notion	🟨	🟦	🟨	🟨	🟦	🟦	🟦	🟦	🟦	🟦

(b) Final design of the toy example for EO



D. *If applicants get same points in the new selection, they should have the same possibility of previous selection or previous rejection.*

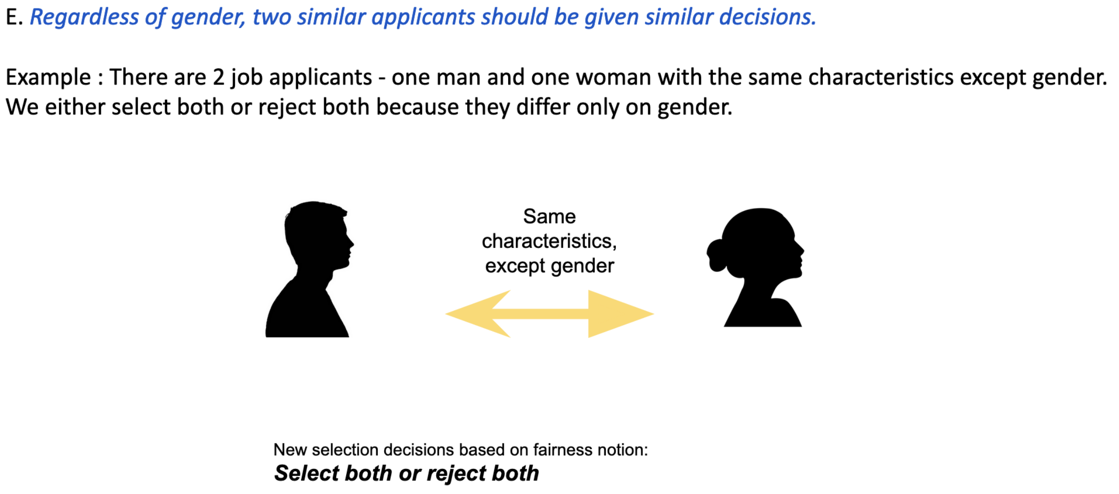
Example: We give the **same** points to applicants who got **similar** decisions before. So, we give **calibrated or relative points** to candidates instead of looking at each candidate individually.

Give points by making use of previous selections.

Consider chance of previous selection or rejection

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Applicant pool of 7 Men and 3 Women										
Previous selection decisions	☺	▬	☺	☺	▬	▬	▬	▬	▬	▬
New points based on fairness notion	9.5	8	7.5	7.5	8	7	7	6.5	6.5	7

(c) Final design of the toy example for CB

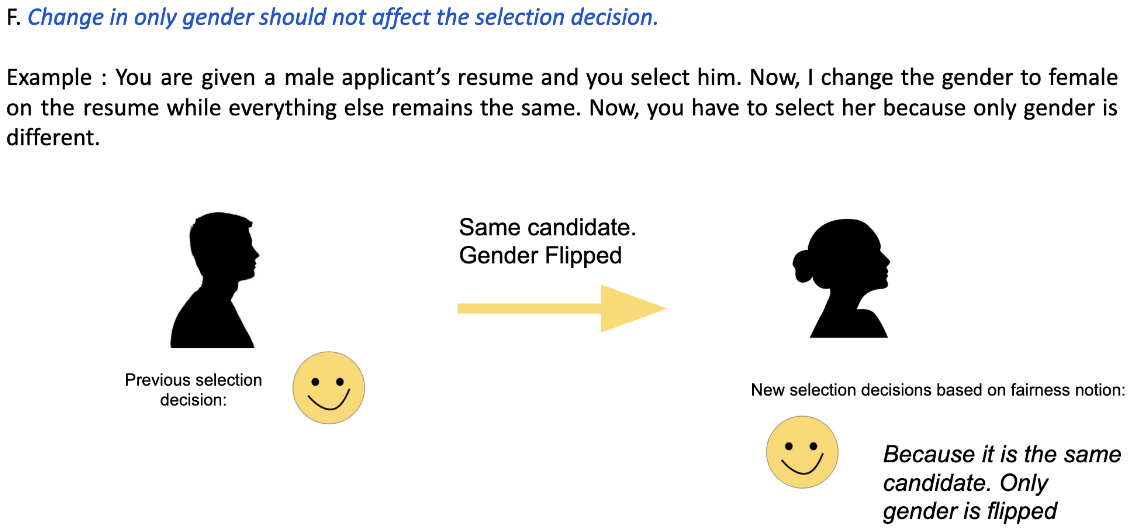


E. *Regardless of gender, two similar applicants should be given similar decisions.*

Example : There are 2 job applicants - one man and one woman with the same characteristics except gender. We either select both or reject both because they differ only on gender.

New selection decisions based on fairness notion:
Select both or reject both

(d) Final design of the toy example for FA



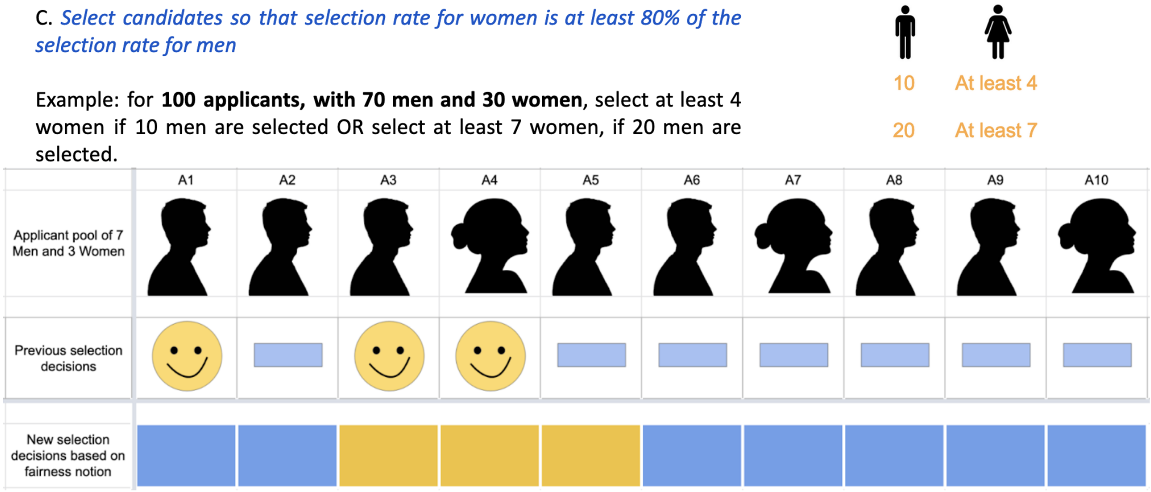
F. *Change in only gender should not affect the selection decision.*

Example : You are given a male applicant's resume and you select him. Now, I change the gender to female on the resume while everything else remains the same. Now, you have to select her because only gender is different.

Same candidate.
Gender Flipped
Previous selection decision:
New selection decisions based on fairness notion:

Because it is the same candidate. Only gender is flipped

(e) Final design of the toy example for CF



C. *Select candidates so that selection rate for women is at least 80% of the selection rate for men*

Example: for **100 applicants, with 70 men and 30 women**, select at least 4 women if 10 men are selected OR select at least 7 women, if 20 men are selected.

10	At least 4
20	At least 7

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Applicant pool of 7 Men and 3 Women										
Previous selection decisions	☺	🟦	☺	☺	🟦	🟦	🟦	🟦	🟦	🟦
New selection decisions based on fairness notion	🟦	🟦	🟨	🟨	🟨	🟦	🟦	🟦	🟦	🟦

(f) Final design of the toy example for DI

C. Paper Submission

The following paper was written during the thesis and submitted to ACM CHI Conference on Human Factors in Computing Systems 2023, an international conference of Human-Computer Interaction (HCI). The submission is currently under review and hence, the authors are shown as anonymous.

“It’s the most fair thing to do, but it doesn’t make any sense”: Perceptions of mathematical fairness notions by hiring professionals

ANONYMOUS AUTHOR(S)

In this work, we explore the alignment of organizational representatives involved in hiring processes with five different, commonly proposed fairness notions. In a qualitative study with 17 organizational professionals, for each notion, we investigate their perception of understandability, fairness, potential to increase diversity, and practical applicability in the context of early candidate selection in hiring. To be more user-friendly to non-technical stakeholders, after initial rounds of interviews with domain representatives, the notions are translated into more accessible illustrations and explanations. Furthermore, we do not explicitly frame our questions as questions of algorithmic fairness, but rather relate them to current human hiring practice. As our findings show, while many notions are well understood, fairness, potential to increase diversity and practical applicability are rated differently, illustrating the importance of understanding the application domain and its nuances, and calling for more interdisciplinary and human-centered research into the perception of mathematical fairness notions.

CCS Concepts: • **Human-centered computing** → **User studies**; • **Software and its engineering** → *Designing software*; • **General and reference** → *Metrics*; • **Social and professional topics** → Computational thinking; • **Applied computing** → Business-IT alignment.

Additional Key Words and Phrases: algorithmic fairness, operationalization, user studies, hiring and early candidate selection, presonnel selection

ACM Reference Format:
Anonymous Author(s). 2018. “It’s the most fair thing to do, but it doesn’t make any sense”: Perceptions of mathematical fairness notions by hiring professionals . 1, 1 (September 2018), 32 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As data-driven decision-making has become prevalent and widely adopted, concerns have been rising about the (un)fairness of algorithmic decisions. With the machine learning components in these decision-making systems typically being guided to optimize for recognizing and reproducing statistical patterns in the data, this risks that majority patterns get amplified and standardized, while minority patterns can be mistaken for irrelevant noise. Furthermore, society has been continuously evolving. Stances on inequality that were considered acceptable a century ago, for example regarding gender or race, may today be explicitly illegal. Still, historical disadvantages and undesired biases still perpetuate into today’s society, and the world as it is today may not yet be the world we want to strive for in the future.

In reaction to concerns that data-driven decision-making can easily scale up and amplify undesired biases, over the past years, dedicated research communities emerged that consider fairness, accountability and transparency, and seek to consolidate legal expertise, expertise from philosophy, and expertise from mathematics and computer science. With regard to fairness, on the side of mathematics and computer science, multiple potential fairness notions have been proposed that can be quantitatively used in data-driven systems. These can be used to monitor the current degree of fairness, or to spark (technical) interventions within a machine learning pipeline to mitigate undesired bias. To ease adoption, several toolkits implementing these notions have been proposed, such as the IBM AI Fairness [4] and FairLearn [8] toolkits.

These notions and libraries may imply that undesired biases in a presently unfair world can be fixed through computational means. The presence of debiasing strategies also is being explicitly used as a selling point by vendors offering data-driven assessment solutions—even though how debiasing will exactly be done tends to remain underspecified [37]. However, the problem is more complicated, as translating real-world problems to data and problem framings compatible with machine learning frameworks is a highly non-trivial matter. The earlier-mentioned question on whether the world generating data and consequent outcomes today actually reflects the world we would like to see (and thus, whether accurately predicting for today's status quo is the goal to strive for, or rather a confirmation of current undesired biases) is key to the way in which a fairness problem should be addressed. The various possible mathematical fairness notions capture fundamentally different world views [35, 20], that mathematically cannot be satisfied at the same time [20, 36]. In technically concretizing what is required to be measured and how optimization should take place (operationalization), both measurement bias and overarching questions of experimental validity need explicit concern, although they are not trivially included in machine learning methodology [24, 33].

In applications of data-driven decision-making in credit risk scoring [23] and criminal recidivism prediction [2], it has been shown that seemingly justifiable choices of operationalization towards increasing fairness actually can perpetuate inequality with strongly adverse effects on the most minoritized populations. Furthermore, with questions of fairness in real-world scenarios typically originating from social science application areas, a potential technical addressing of fairness (which typically will be built by computer or data scientists) needs to be contextualized at interdisciplinary crossroads. Here, the computationally and non-computationally-minded stakeholders may think they speak of the same problem, but actually depart from different underlying assumptions on which aspects of the problem need the deeper research [33].

While the discussion so far focuses on technical research and innovation, problems of (un)fairness also are actively being addressed as an organizational operational concern, as e.g. exemplified by the establishment of Equity, Diversity and Inclusion offices and interest groups, that seek to implement fairness-promoting strategies in the policies of organizations. Stakeholders involved in these actions typically do not fundamentally perform academic research on matters of fairness, and thus are not trivially engaged with advances proposed in the field. As such, to avoid that a research-practitioner gap may occur, as happened in personnel selection [33], it is worthwhile to try engaging these audiences.

In this work, we seek to do this, by exploring the alignment of organizational representatives involved in hiring processes on five different, commonly proposed mathematical fairness notions. Potential advantages of these fairness notions are that they are explicit with clear boundary definitions, and as such may be a way to very crisply define policy. However, as potential disadvantages, they may take a very simplified and unimplementable take on the hiring and selection process, and not be understandable to professionals who may not be mathematically inclined.

To this end, after surveying current practical experiences on implementing fairness-promoting policies, we design more accessible and user-friendly explanations of the chosen fairness notions, and use them to initiate dialogues with the participants on the extent to which these notions are understandable, fair, and applicable in practice. More concretely, we depart from the following main research question and sub-questions:

How do organizational representatives understand and perceive different mathematical fairness notions in the context of early candidate selection in hiring?

(1) How do organizational representatives align with different mathematical fairness notions in the context of early candidate selection in hiring?
   (a) What is their conceptual understanding of the mathematical fairness notions?
   (b) What do they rate the mathematical fairness notions (and their consequent output decisions) in terms of fairness and why?
   (c) What do they rate the mathematical fairness notions in terms of improving diversity in organizations?

(2) What is their impression, perception of fairness, perception of diversity and applicability of using the different mathematical fairness notions during the early candidate selection in hiring?

We will investigate these questions in a qualitative study, based on semi-structured interviews with professionals in executive functions, talent acquisition, HR, I/O psychology, and diversity and inclusion operations.

With this study, we offer several contributions to the field. First of all, there still is much less empirical work on human perceptions and considerations of technical fairness notions and potential interventions, than there are proposals of new notions or interventions. Existing work focusing on human perception of fairness tends to focus on lay people's understanding of generic fairness scenarios [39], [38], [22], [30]. In contrast, in our work, we address participants who are already professionally committed to addressing questions of fairness, albeit from a very different methodological angle than that of the algorithmic fairness domain.

In focusing on hiring as an application domain, we choose a scenario that has high-stakes elements, and that is compatible with Equity, Diversity and Inclusion special interest work in organizations. At the same time, we take a different focus from common research on algorithmic fairness in hiring, which would normally focus on AI-based automated assessments of job candidates [33, 37]. In contrast, we choose a specific point in the hiring pipeline: that of early selection, in which an initial large group of potential applicants needs to be reduced to a smaller group, that will advance to further assessment or interviewing rounds. By focusing on this point in the pipeline, we deliberately do not look at the highest-stake phase in the hiring processes where actual offers will be made with a final candidate. The early selection phase also is naturally compatible with diversity-improving interventions in the real world: if an organization is unsure but curious about a candidate, even while other candidates more obviously seem a match, the candidate can still advance. Towards our participants, we will actually leave ambiguous whether algorithmic processes or humans would make a selection; instead, we primarily want to focus on the extent to which very formalized policy notions may be compatible with current human selection practices.

Finally, in conducting this study, we have been raising awareness across fields on current best practices. With the authors of this work being computer scientists, the study helped in gaining deeper insight on how to navigate requirements on fairness-promoting interventions with domain experts. At the same time, for many of the domain experts, this was the first time to be familiarized with algorithmic and more formalized takes on fairness concepts.

2 RELATED WORK

2.1 Candidate selection processes

Organizations manage operations related to recruitment, hiring, employee satisfaction, promotion and retention through Human Resources Management (HRM), making it an important source of information to design informed policies concerning their employees. In recent years, these functions have expanded to include diversity and inclusion operations within organizations. These can start from sourcing potential employees to finally integrating new and existing employees in diverse set-ups. Hiring is the first stage accessible to HRM functions to work towards this goal.

The hiring pipeline of organizations can be broadly composed of three stages: recruitment, selection and job offering. Recruitment focuses on targeting and attracting potential employees for different job vacancies, with the goal of receiving job applications from interested applicants. The second stage, candidate selection, assesses and evaluates the job applicants through multiple steps of screening, tests and interviews. In these steps, applicants are evaluated by multiple decision-makers such as Human Resources (HR) or vacancy-creators, thereby producing a smaller subset of applicants with each step. Among the final subset, the most suitable applicants are offered the jobs.

The candidate selection stage has seen an increasing adoption of data-driven algorithmic decision-making interventions offering automated candidate screening and assessment [14, 37, 10, 17]. Reasons for the adoption can be attributed to overcoming unconscious human judgment [27], increasing efficiency in processing [9], and economic benefit [1]. However, data-driven algorithmic decisions have been shown to reinforce human bias [37, 13]. A study [32] found that while some organizations onboard automated candidate screening and assessment tools after legal consultation, only a few consult experts on the input features, data, validation methodology or fairness notions used. Without fully understanding the implications of the tool, the adoption of automated systems limits HRM from justifying its use and subsequently, achieving their goals of fair and inclusive hiring.

2.2 Algorithmic fairness

Group Fairness, also known as statistical fairness, seeks to treat different social groups equally [29]. After fixing a number of groups based on sensitive features, some chosen statistical measure needs to be valid across all the multiple groups [16]. Group fairness is easy to understand and obtain as it does not require any assumptions of the underlying population [26]. This level of analysis provides average guarantee for groups. One implementation checked the average weighted similarity of outcomes for two groups [5]. However, the average guarantees do not highlight the instances where some people are still unfavorably treated [16]. Moreover, some statistical fairness notions can be satisfied without reaching the social goals of fairness [26] and thereby, fail to provide meaningful guarantees [19]. Studies also show that multiple statistical notions are impossible to satisfy simultaneously [28, 15], making it difficult to operationalize group fairness.

In contrast, individual fairness asks for similar predictions for similar individuals [29]. A fine-grain analysis of fairness can be achieved by placing constraints on pairs of individuals [19]. A motivation for desiring this granularity of fairness is that people who are less qualified should not be preferred over more qualified ones [25]. Moreover, individual fairness provides better understanding of the social context than group fairness [16]. One implementation for satisfying fairness at individual level compares every individual of a protected group to every individual from another protected group and penalizes the algorithm for different decisions for similar individuals [5]. A distance or similarity metric is required to compare individuals to see how similar they are. The downside to this is that different metrics can produce different results. Therefore, knowledge of the data, especially, the relationship between features and labels are required to perform such fine-grain analysis [16], [26]. This also requires domain knowledge and understanding of organizational requirements.

2.3 Human-centered research on fairness notions

Several studies have been conducted to understand lay people's perception of fairness of data driven algorithmic decisions. On surveying lay people, it was find that people tend to believe statistical parity as fair because of its simplicity, while other complicated notions are deemed unfair [39] in recidivism and cancer prediction. Moreover, [38] studied three notions - fairness through awareness, calibrated fairness and never favor a worse individual over better one. By presenting scenarios in the domain of lending, the research investigated lay people's perception of fairness and how it changes when an additional sensitive feature is presented. They conclude that calibrated fairness was the most preferred notion in the context of lending.

Another research [22] studied how people perceive the usage of different input features, including sensitive features for making recidivism prediction decisions. They found that there is no consensus in the fairness of input features used, unfairness concerns of people go beyond discrimination and features can be considered unfair to be used despite not being a sensitive feature such as education, quality of life or involvement with family. Lastly, to understand the fairness judgment and the emotions evoked, [30] surveyed people in the context of managerial task to find that people consider algorithmic decisions as fair for mechanical tasks such as assigning work or scheduling, whereas they think algorithmic decisions are less fair than humans in tasks such as selecting candidates for a job or evaluating performances. They conclude that people's knowledge of the decision makers has an influence on the validity and trust in the decision made.

Moving from ordinary people's perception of fairness, [42] and [12] investigate how people affected by algorithmic decision feel through face to face interaction with people. The research from [42] interviewed people from marginalized communities and found that algorithmic outcomes evoked negative emotions regarding racial and economic injustice and reduced their trust of the products of companies using algorithmic decision making. Furthermore, the research from [12] interviewed multiple stakeholders in child welfare service and found that people have many concerns with the system, the different scenarios in child welfare service and the degree of reliance on algorithmic decisions. The research also highlights affected people's concerns with bias in data and biased decisions made by the human relying on algorithmic decisions. Lastly, [31] uses participatory design by interviewing multiple stakeholders to obtain best practices to design an algorithmic decision making service to allocate donation to non-profit organizations. They conclude that different stakeholders have different fairness motivations, which depends on people, context and the type of interaction they will have with such systems.

The effect of explanation styles for algorithmic decisions have also seen some work. By surveying people on 5 common situations through fictional scenarios such as loan, work promotion, car insurance, flight re-routing and freezing of bank account, [7] conclude that there is no best way to explain algorithmic decisions to people. The reason being that explanation styles does not effect how people respond to (in)justice given to the fictional scenarios in the 5 situations. People's concerns with algorithmic decision making include lack of human involvement, reasonable outcome and morality. However, the study [18] shows that people consider some explanation styles unfair while some explanation styles increase their trust of the system in the context of recidivism prediction, where the severity of the decision is higher than the common situations in [7]. Moreover, people's prior position on algorithmic fairness has an influence on their reaction to algorithmic decisions.

3 INITIAL DOMAIN INTERVIEWS

To better familiarize ourselves with our chosen problem domain, understand the candidate selection process and current ambitions of HRM professionals regarding questions of fairness, and make more informed choices on the scoping of our main study, semi-structured interviews were first conducted with a small group of organizational professionals. Participants were recruited by direct e-mails obtained from connections of the research group. A total of 5 professionals participated in the semi-structured interviews, out of whom 4 are female and 1 male, with working experience in the field ranging from 0.5 to 23 years. All the 5 participants work in [ANONYMYZED COUNTRY] and none of the participants or their organizations have used automated decision-making in the context of early candidate selection in hiring. The participant information is shown in Table 1.

Table 1. Background information of the participants for initial domain interviews.

Participant ID	Job Title	Experience	Gender
PS1	Professor, Psychology in hiring practices	20 years	Female
PS2	HR Advisor	5 years	Female
PS3	Senior Talent Acquisition	23 years	Male
PS4	Talent Acquisition specialist	4 years	Female
PS5	Admission Coordinator, Bachelor Programme	0.5 year	Female

Each participant was interviewed independently for one hour. The participants' informed consent was obtained by presenting them with an opening statement containing the goal of the interview, information to be collected (the information stated in Table 1, and the contact details of the researchers. The semi-structured interviews covered 4 broad topics as follows:

- How does the current candidate selection process work?
- What goals do the different stakeholders strive for during recruitment and early stages of candidate selection?
- How is fairness, diversity and inclusion handled in the current selection process?
- How can fair selections be made in the early stages?

The responses were meant to inform our study design, but were not intended to serve as formal analysis data of our study. Therefore, written notes were taken during the interviews, but no recordings or transcriptions were made, to more strongly protect the privacy and anonymity of our participants.

3.1 Findings

3.1.1 How does the current candidate selection process work? The early selection process is performed by the vacancy-holders, who create the job vacancies and selection committees appointed by HR functions. The main role of the HR functions is to ensure transparency of the process. The selection-makers have the option to explain their choices and decisions about each candidate they review. However, they are not required to provide an explanation. Most selection-makers make notes about candidates. According to the participants, most organizations do not have well formulated or explicit diversity criteria during selections.

3.1.2 What goals do the different stakeholders strive for during recruitment and early stages of candidate selection? The goals for different stakeholders such as the organization, HR functions and selection-makers differ. The organization looks for candidates who match the organization's overall organizational goals, make the organization's population diverse and enable the company's growth, whereas HR functions help find a diverse applicant pool for the different job vacancies in the organization. The utility for selection-makers comes down to finding candidates with the right skill set, motivation and capability to do the job well.

3.1.3 How is fairness, diversity and inclusion handled in the current selection process? Organizations take various measures to operationalize fairness and diversity in the selection process. They begin by creating textual descriptions of job vacancies that are likely to attract a diverse applicant pool. Internally, organizations may also provide bias and non-discrimination training to the selection-makers and HR functions. Legal compliance is also taken into consideration.

Many organizations also have diversity in their statements of mission, which may or may not be known by selection-makers. Organizations are allowed to access some sensitive features of the applicants. Some use gender as a factor while making early selections, but do not have policies or directions on monitoring other sensitive features.

3.1.4 How can fair selections be made in the early stages? All participants identified gender as an important sensitive feature, while some also indicated nationality and ethnicity as important. However, none were able to coherently argue about the need to monitor different sensitive features—we asked this question explicitly, as it will define what aspects of fairness will be noticeable at all in data. The confusion was caused by legal limitations and company policies on the use and access of the sensitive features. Further, every participant stated that it was impossible to monitor False Negatives (i.e., rejected candidates that should in retrospect have been selected) in the selection process as per the GDPR (General Data Protection Regulation) guidelines, because applicant information cannot be retained by the organization without the applicant's consent. However, all participants indicated the possibility of monitoring for False Positives (i.e., accepted candidates that should in retrospect not have been selected) in the selection in the future.

3.1.5 Consequences on our study design. The responses to our questions helped in refining the design of our main study. First of all, it was clear that in recruiting future expert participants, we should strive for organizational representatives who have concrete involvement and influence over the early selections in hiring. As for characteristics on which fairness examples could be based, gender was the only sensitive attribute that today may more explicitly be considered and registered. Thus, in considering potential adoption of mathematical fairness notions as part of diversity-promoting policy (or, as an extension of this, fair machine learning interventions), this may be the only sensitive attribute that could materialize as data on which the fairness notions could realistically be applied. Therefore, we chose to design explanations of mathematical fairness notions that would deal with situations in which there is a skewed balance between men and women.

Among our five participants, we included an admission coordinator for a university's bachelor programme. This was because we wondered whether student admissions could be seen as a similar problem to early candidate selection. While this participant's answers on the student admission procedure had many similarities to the answers of HR professionals (e.g., 'matching organizational goals' would in this case be 'being likely to achieve study success'), the procedure to go from an initial set of applicants to a smaller set of selected students was much more strongly driven by standardized tests than in job hiring cases. As a consequence, for our main study, we chose to really focus on early selection of job candidates.

4 DESIGN OF EXAMPLES FOR THE MAIN STUDY

In our main study, we make use of examples of early candidate selections based on various mathematical fairness notions, which we want to translate in a non-mathematical way. After discussing which mathematical fairness notions we chose to adopt in our study, we discuss their translation into more accessible, non-technical example illustrations.

4.1 Chosen mathematical fairness notions

For our study, we adopted 5 different fairness notions (3 group fairness notions and 2 individual fairness notions), which both are well-known in literature on algorithmic fairness, and realistically applicable to early selection stages.

Formally, all of the notions can be implemented as optimization constraints in a binary supervised machine learning classification problem. In this problem, we have a collection of N candidates, where each candidate is represented by a given feature set X, a single binary sensitive feature, A and true outcome, Y, which is a binary variable indicating whether a candidate was selected to advance to the next stage or not. A classifier makes predictions $\hat{Y}$, that should be as close as possible to Y, while the fairness notion of interest is being satisfied.

4.1.1 Statistical Parity. Statistical parity (SP) is a group fairness notion, that requires for the prediction $\hat{Y}$ to be statistically independent of the sensitive attribute $A : P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$. This means that equal prediction rates across groups should be reached, regardless of the actual outcome Y [19]. SP is a suitable metric when there are legal requirements of equal acceptance rates for multiple sensitive groups. However, the downside is that it can be satisfied without satisfying fairness. For instance, to obtain equal acceptance rates for men and women in hiring, the recruiters can select qualified candidates from one group whereas, select only random candidates from the sensitive group to satisfy this fairness criteria. This results in masking [3], where random candidates instead of qualified candidates are selected in order to satisfy statistical parity.

4.1.2 Equal Opportunity. Equal Opportunity (EO) is a group fairness notion, where the positive prediction $\hat{Y}$ should be conditionally independent of the sensitive feature, given that Y comes from the positive class : $P(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1|Y = 1, A = 1)$. This means that the probability of being predicted in the positive class when the actual outcome is positive should not depend on the sensitive feature [23]. An example satisfying equal opportunity is when equal proportions of people are selected from the qualified fraction of the sensitive group, such as men and women. This fairness notion is useful when the False Positive (FP) rate is not important. In practice this would mean that more unqualified employees also get selected for the next round in the hiring process, along with the qualified employees. This can be considered fair because it gives equal opportunity to all candidates, irrespective of the sensitive feature. This also implies that equal opportunity should not be applied when having high FP can have consequences (e.g when firing ill-performing employees, many well-performing employees would also get fired) [34].

4.1.3 Calibration. Calibration (CB) is a group fairness notion, requiring equal probability of belonging to the positive class for the same predicted score $S = s$, irrespective of the sensitive feature : $P(Y = 1|S = s, A = 0) = P(Y = 1|S = s, A = 1)\ \forall s \in [0, 1]$ [15]. For example, for men and women with a predicted qualification score of 0.8, there should be equal probability that their actual outcome was positive. The same statement holds for every value of $s \in [0, 1]$.

4.1.4 Fairness through awareness. Fairness through awareness (FA) requires the same prediction for any pair of individuals whose similarity falls under a given threshold [19]. For any two individuals i and j, where $P(X_i)$ is the probability distribution over all possible outcomes of prediction for i, D measures the distance between two probability distributions and d measures the similarity distance between the two individuals, fairness through awareness can be written as : $D(P(X_i), P(X_j)) \leq d(X_i, X_j)$. For instance, for a binary outcome, if the probability distribution for individuals i and j are [0.3,0.7] and [0.2,0.8], respectively, the distance between the distributions could be measured by, say Hellinger distance between them which is approximately 0.08. Here [0.3,0.7] means that the probability of belonging to the positive class is 0.3 and the probability of belonging to the negative class is 0.7. Now, if the similarity distance metric d between i and j is, say the euclidean distance between them, fairness through awareness is achieved if the Hellinger distance of 0.08 is lower than the euclidean distance between features of i and j. FA provides fine-grained analysis, because it can quantify how individuals are treated. However, a major challenge of this fairness notion is that a good collaboration is required between domain experts to define the similarity metric, which can be different based on the context and its requirements.

4.1.5 Counterfactual Fairness. Counterfactual fairness (CF) is an individual fairness notion, that requires the probability for every individual, i with $A = a$ to get the same prediction, had the sensitive value been $A = a' : P(\hat{Y}_i|A_i = 0) = P(\hat{Y}_i|A_i = 1)$. . It looks at the causal relationship between variables, rather than the statistical correlations between them [29]. For example, looking at the change in probability of receiving a positive or negative outcome for an individual with her ethnicity flipped, can show how the model is dependent on ethnicity for the prediction, indicating the extent of unfairness according to CF?

4.2 From mathematical notions to accessible non-technical examples

With regard to the design of accessible explanation examples, an iterative process was followed. An initial design, as shown in Figure 1a, was visually based on a similar design used in [39] for testing lay people's perceptions of fairness in recidivism predictions. Textual human explanations of the mathematical notions were drafted by the authors. The design was discussed with a lab member independent of our current study. This lab member has hands-on background in developing social inclusion programmes in the cultural domain, is starting to get familiarized with concepts of responsible AI applications, but explicitly does not have a technical or mathematical background.

Following multiple discussions with this lab member, as well as between the involved authors, several design changes were made. First of all, the framing of an 'algorithmic prediction' seemed to distract from the true purpose of the study: assessing to what extent common algorithmic notions match current practitioner thinking about selection policies, regardless of whether human or algorithms are the entity ultimately making the selection. Therefore, we ultimately removed any reference to algorithmic procedures. Instead, the fairness notion of interest would be contrasted with 'what a previous selection committee' would have judged. In making this choice, we also avoided potential response bias, as the type of decision-maker for the same task evokes different emotions in people [30]

Next to this, we found the concept of 'true outcome' problematic in the context of potentially biased hiring procedures: it could imply an absolute ground truth that a prediction mechanism should match as closely as possible (which indeed is the framing commonly seen in machine learning). To keep the option open that previously known judgments may not necessarily be repeated, we framed the previous judgments as judgments from 'a previous selection committee'. We did not state whether this committee would have been undesiredly biased, but left this up to the interpretation of our study participants.

In a similar line of thought, the concepts of 'selected' and 'rejected' were deemed too harsh: it may imply that those rejected would be unqualified or inferior to those selected, where it actually is unknown whether this truly would be the case. For subjects who historically would be selected, we know they positively stood out to the party doing the selection, but for those who would not, there actually is no information on whether these candidates actually are worse, and as argued in Section 3, legal restrictions currently make it impossible for this to be verified.

Therefore, we chose to visually give a positive association to those who would previously get selected, but give a more neutral indication for those who would not. Generally spoken, the choice of a red vs. green color palette would not be colorblind-friendly, while at the same time again evoking stronger 'good' vs. 'bad' associations than we wished to imply. Instead, we chose to go forward with a more accessible blue vs. green palette.

Finally, where we initially chose a slightly disbalanced gender sample (6 men and 4 women) where fewer or more people could be selected, we amplified the disbalance to 7 men vs. 3 women, out of which only a fixed, smaller number could be selected. In doing this, we wished to have the disbalance standing out more, while at the same time being firmer on limited slots being available in the fictional next application stage. The final design is shown in Figure 1b

In parallel to the toy example illustrations and their choices of framing and wording, we also discussed how understandable our textual human explanations of the mathematical notions were, and whether some notions may be easier or more difficult to grasp, potentially affecting whether they should be presented towards the beginning or end of an interview.

Our initial textual explanation was still deemed too mathematical, and needed several rounds of explanation for each notion. Upon understanding the concepts, for the group fairness notions the lab member was able to reflect on SP and later compare it with EO. She said she needed to understand SP before she could comment on EO, as EO was an 'improvement' on SP. This remark was consistent with the finding in [39], where lay people selected SP as fair, but had difficulty judging the other, more complicated fairness notions. The lab member found CB most difficult to understand. As for the individual fairness notions, CF raised questions from the lab member about change in gender identity of the applicant and its consequences. FA triggered less additional discussion, and thus would be easier to discuss first.

Following our refinements to our example designs and explanations, we tested these with two further independent lab members with technical machine learning backgrounds. They were first shown our designs, after which they were shown the mathematical definition for comparison. With their suggestions on further clarifications, we finalized our designs. The final full set of designs is presented in Appendix A.

5 METHODOLOGICAL SETUP

5.1 Responsible research practices

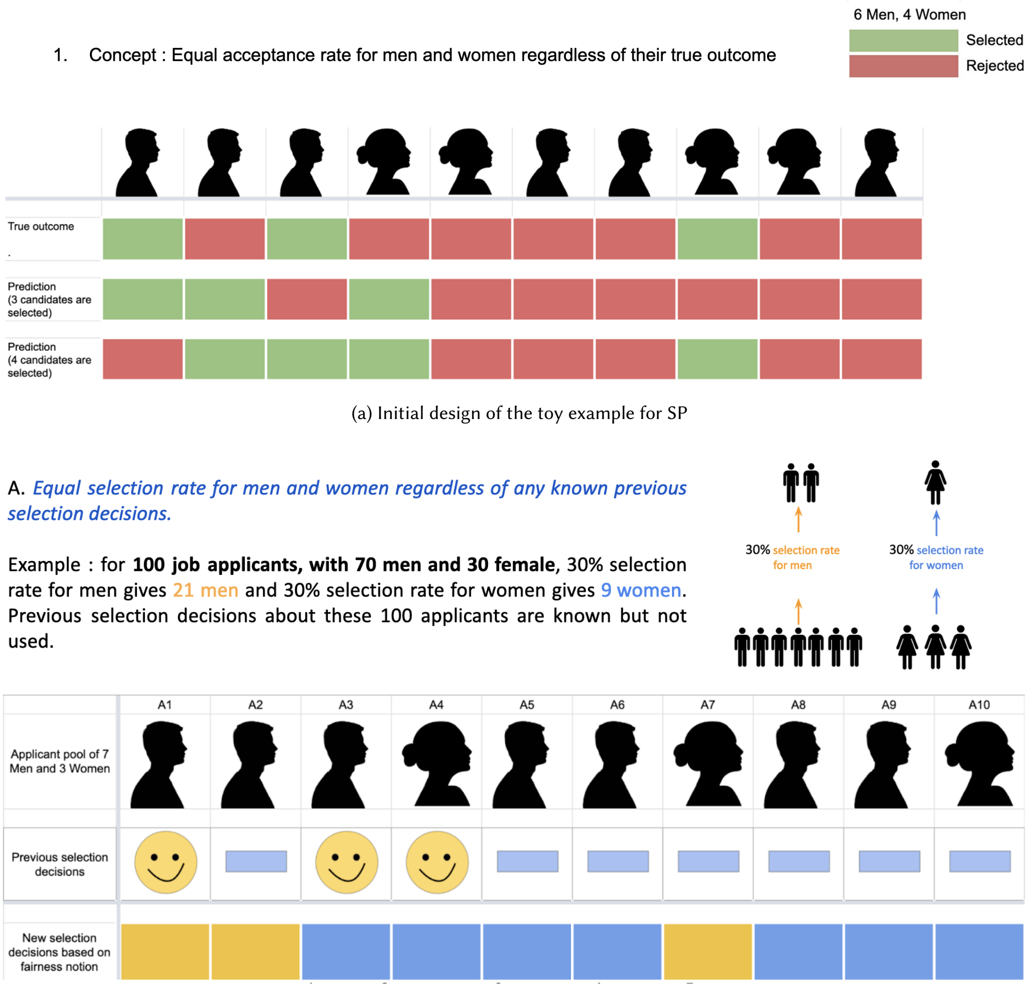
With our study considering research with human subjects, Human Research Ethics Approval was requested and granted at the main authors' institution. Obtaining human research ethics approval also required for the opening statement to be externally reviewed, and a separate Data Management Plan to be written and approved. For the sake of transparency and reproducibility, we release our codebook and corresponding participant quotes (following explicit participant consent) as supplemental material. However, to respect the privacy of our participants, full audio recordings or transcriptions will not be reshared.

5.2 Recruitment

We used a combination of purposive sampling and snowball sampling [21] to recruit experts. Initially, we sent direct e-mails to relevant professionals in the domains of HRM and Diversity and Inclusion (D&I) operations found with the help of our network of our personal connections in industry. Next, we made announcements on LinkedIn inviting professionals in the same domains to participate in an interview by sending us their e-mail. We also found a list of professionals working in this field using LinkedIn, whom we contacted by sending direct messages. Overall, we contacted over 80 potential participants, out of whom 48 responded, and 21 scheduled interviews. We conducted the study at [ANONYMIZED INSTITUTE].

5.3 Participants

Out of the 21 participants, we selected 17 for our more thorough analysis, who are most relevant to the field of this study and have experience with early selection processes. Their professional role was checked with their Linkedin profile. The demographic information of the 17 participants is listed in Table 2. Out of the 17 participants, 5 self-reported as belonging to an ethnic minority, 4 self-reported as belonging to other minority related to age, health, sexual orientation,



1. Concept : Equal acceptance rate for men and women regardless of their true outcome

True outcome	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟥	🟥
Prediction (3 candidates are selected)	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥
Prediction (4 candidates are selected)	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟥	🟥

(a) Initial design of the toy example for SP

A. *Equal selection rate for men and women regardless of any known previous selection decisions.*

Example : for **100 job applicants, with 70 men and 30 female**, 30% selection rate for men gives 21 men and 30% selection rate for women gives 9 women. Previous selection decisions about these 100 applicants are known but not used.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Applicant pool of 7 Men and 3 Women										
Previous selection decisions	🙂	▭	🙂	🙂	▭	▭	▭	▭	▭	▭
New selection decisions based on fairness notion	🟨	🟨	🟦	🟦	🟦	🟦	🟨	🟦	🟦	🟦

(b) Final design of the toy example for SP

Fig. 1. Design changes to the toy example

immigration and neurodiversity, and only 1 belonged to both ethnic and another minority group. All participants work in [ANONYMIZED COUNTRY], at a diverse set of differently-sized organizations, as shown in Table 3.

5.4 Interviews

The same researcher interviewed each participant independently, either in person or online using Microsoft Teams. The semi-structured interviews were conducted in one or two sittings, which overall lasted 75 minutes on average. The interview began by presenting an opening statement containing goal of the interview, participant information to

Table 2. Demographic Information of Study Participants.

	Job Title	Experience	Education	Gender	Ethnic Minority*	Other minority**
E1	Talent Acquisition Specialist	23 years	Bachelor	Male	No	No
E2	D&I Officer	4 years	Doctorate	Female	No	No
E3	HR Manager	24 years	Master	Female	No	No
E4	Managing Director	3 years	Master	Male	No	No
E5	Executive Board Member	5 years	Bachelor	Female	No	Yes
E6	CEO	14 years	Master	Male	Yes	No
E7	HR Development Trainee	1 year	Master	Male	No	No
E8	HR Business Partner	4 year	Master	Male	Yes	No
E9	HR advisor	5 years	Doctorate	Female	No	No
E10	D&I Advisor	8 years	Master	Female	No	Yes
E11	Chief Diversity Officer	10 years	Doctorate	Male	Yes	No
E12	Recruitment Technology Consultant	15 years	Master	Male	No	No
E13	Assistant Professor (as vacancy holder)	6 years	Doctorate	Female	No	No
E14	Psychological Assessment Reseacher	40 years	Master	Female	No	No
E15	Global D&I Manager	12 years	Master	Female	Yes	Yes
E16	Inclusion Specialist	6 years	Bachelor	Male	No	Yes
E17	I/O Psychologist	7 years	Master	Female	Yes	No

Ethnic Minority* = self-reported as belonging to an ethnic minority
Other Minority** = self-reported as belonging to a minority group that faces discrimination

Table 3. Participant distribution over organization industries and sizes.

Employment Industry	Number of Participants	Organization Size
Education and Sports	1	<10
Recruitment Service	2	<10
Water Management	1	<100
Human Rights	1	<100
Arts and Culture	1	<100
Higher Education and Research	3	<1,000
Higher Education and Research	4	<10,000
FMCG	1	<70,000
Police	2	<70,000
Offshoring	1	<70,000

be collected, purpose of audio recording, participant's right to withdraw from the research and contact details of the researchers. After obtaining the participant's explicit and voluntary consent to the opening statement, we proceeded with the semi-structured interviews.

The interview consisted of 5 parts:

(1) Introductions were made and the participants provided their background information, shown in Table 2.
(2) The participants were introduced to the context of early candidate selection, example job positions and multiple vacancies.
(3) For each fairness notion, participants were explained the notion according to our non-technical translation, and were asked to think out loud about their understanding, concerns, benefits, feelings and implications of adopting this fairness notion. After participants described their initial thoughts, they were shown the pictorial representation of the same fairness notion and their comments were obtained. The discussion of each fairness

notion concluded with 5-point Likert scale responses on understanding, fairness and diversity in terms of gender, which is describe in the following questions:

- What would you rate your understanding of this fairness notion?
  (1=Don't Understand, 2=Somewhat Don't Understand, 3=Don't know, 4=Somewhat Understand, 5=Understand)
- What would you rate this notion on fairness?
  (1=Unfair, 2=Somewhat Unfair , 3=Don't know, 4=Somewhat Fair, 5=Fair)
- What would you rate this notion's ability to improve gender diversity?
  (1=Unhelpful, 2=Somewhat Unhelpful , 3=Don't know, 4=Somewhat Helpful, 5=Helpful)

In terms of the order of presentation, we randomly started with the set of group or individual fairness notions. Within these sets, we retained the same order of presentation (group: SP, EO, CB; individual: FA, CF), which would be in line with increasing difficulty as perceived in our design phase, described in Section 3.

(4) After all fairness notions were presented, the participants were asked to re-rate each of the notions again on the same 5-point Likert scale. In doing this, we wanted to see whether a re-evaluation after seeing all notions would lead to changed ratings.

(5) Finally, participants were asked to describe their experience of the interview and their thoughts about using the fairness notions.

5.5 Analysis

Audio recordings of the interviews (totalling 26 hours of interview content) were automatically transcribed through the automatic closed captioning functionality offered in Adobe Premiere Pro 2022 for audio interviews and Microsoft Teams for video interviews. These transcriptions were manually checked by the two authors for correctness of translation with the help of the corresponding audio or video recordings and later imported into the Atlas.ti software, which was employed to conduct a thematic analysis [11].

Initially, the two authors independently coded 4 interviews employing an inductive approach, having codes of relevance emerging from the text. Then, following the main research questions, but considering these next to the codes obtained through the inductive analysis, a reference codebook was established, and a deductive approach was followed by one of the authors to code all 17 interviews, consolidating style and granularity of the two initial coders. The codes were categorized into high-level themes and grouped per fairness notion.

6 RESULTS

In reporting our results, we visualize the ratings given by participants on Understanding, Fairness and Diversity. Beyond an overall visualization, we also show breakdowns by sensitive self-reported features (Gender, belonging to an Ethnic Minority, belonging to an Other Minority). While with 17 participants, we quantitatively cannot make strong statistical claims, this can help visualizing whether participant's sensitive features may potentially lead to different rating behavior. Following our thematic analysis on the interview transcripts, for each fairness notion, we also qualitatively discuss responses to impression, perception of fairness, perception on improving diversity, and applicability. Lastly, we discuss the outcomes of the renewed ratings on fairness, that participants did after seeing all the fairness notions.

6.1 Statistical Parity (SP)

We present the qualitative analysis for SP from 150 quotations obtained upon coding transcripts of 17 participants. Rating distributions of the participants are shown in Figure 2.

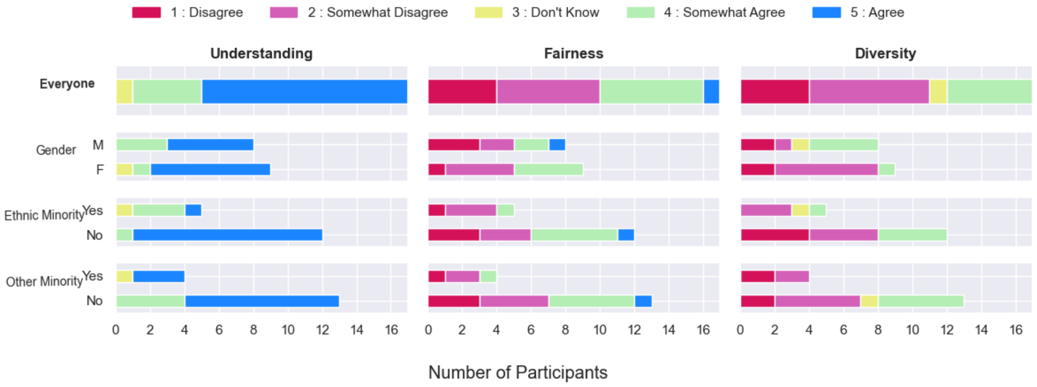


Fig. 2. **SP** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of SP, along with the distribution by *gender, ethnic minority* and *other minority* groups.

6.1.1 Impression of SP. SP was well-understood by all participants. At the same time, participants had a wide range of thoughts and comments on the notion. Six participants expressed what they liked such as merit being the basis of selection (E1, E9, E12) and that the notion objectively treats all job applicants equally (E2, E4, E6). *"..it should not restrict us getting in contact with good candidates"* - E9. "creates equal opportunities for both sexes, which is positive" - E6. However, two of the same and four additional participants also disagreed with the notion or said that they would not use the notion in their organizations (E5, E6, E8, E11, E12, E16). "I understand what it's trying to do. But would I use it? Do I agree with the fairness notion, then no." - E5. "I would never do it this way. I don't care what the gender is in this stage. I want to select the best candidates." - E12. In fact, several participants expressed their dislike for the notion directly or indirectly. While some participants directly mentioned that they did not like the notion (E5, E6, E7), others did not like gender being the basis of separation (E4, E1, E7). "I don't like this one. It is very well possible that these 70 men aren't very qualified at all and then you're going to still hire 30% of them" - E7. " ..you are making a difference on gender and it's something we try to avoid as long as possible." - E4.

While one participant was glad that Statistical Parity allowed some representation of minority groups, which might not otherwise happen without effort (E10), three participants said that the notion favors the majority and may not help minorities (E2, E10, E15). "I'm happy that there's at least more than just one woman, because we often see there's only one woman. Looking at chances, the men have more chance of being hired than the women do in this case." - E10. "We know that the chance of hiring female is zero. Right?" (sighs) - E11.

6.1.2 Perception on fairness of using SP. There was no consensus in the ratings, with participants using the full Likert scale in their responses. A similar pattern of ratings is seen for both genders and majority groups.

About nine participants, argued this notion as unfair because they feel it is not equitable (E8), is based on gender (E5, E6), is based on quota systems (E11), skewed applicant pools (E3, E15, E16) and other factors (E1, E2). "it's not equitable.

So I think it's unfair". - E8. "The whole system is not fair because you don't want to use a quota system, right?" - E11. On further examination, we see that many participants consider SP as fair and applicable only in an ideal world (E2, E4, E8, E9, E13, E15, E16, E17). "It's fair in theory, unfair in result. If it were a fair world, a perfect world, then this would be a fair procedure." - E2. According to them, theoretical fairness is attributed to ideally treating everybody equally. "Technically, it's still fair because of same selection rate for men and women" - E4. However, seven of them hesitated about the calling the notion fair. "I think it's fair on paper [...] it's actually not quite fair because 30 women and 70 men applied for the job, and it feels really wrong to have only 9 women go on to the next round" - E16. Lastly, four participants found it quite difficult to say whether or not this notion could be called fair. " this is a difficult one. I really would like to know why A1, A3 and A4 were chosen by the former committee and why the decision is now different. If I don't know why, I cannot say if it's fair or not, or sensible." - E14.

6.1.3 Perception on improving diversity when using SP. Participants were more certain about commenting on improving gender balance in organizations with the help of Statistical Parity. Five participants felt positive about the notion's ability to help diversity in terms of gender (E1, E2, E4, E7, E9). "We have 75% men and 25% women in our company. So when we introduce this fairness notion, then it will help to balance our company more." - E4. They liked the notion, and reasoned that despite a skewed applicant pool, it can act as a precautionary step because final hiring can be biased, however long the improvement takes. "You will provide possibility of balance in your selection group but then there's still the final selection". - E1. "But if you really want to hit the targets, that's probably not going to help." - E7. This can be seen in the ratings, where seven participants gave a rating of 4 or higher for diversity.

However, a larger number, totaling eight participants, reasoned about the notion's inability to improve diversity for two main reasons (E2, E3, E5, E8, E12, E13, E15, E16). First, four participants said that diversity goals cannot be achieved if the minority is absent in the applicant pool. "Because if only one female applies, then there goes your theory" - E5. Second, six of them said that broader diversity goals of the organization cannot be realized because of small minority representation in selections. "You need to identify your minority group and make sure that's at least 50% of your talent pool. Otherwise you'll never make this change." - E8. Despite, polarized perceptions, five participants indicated that the notion could contribute to diversity to some extent (E1, E4, E6, E11, E13). "I mean of course, if your team had no women then you're improving your gender balance. If your team consisted of only women, you might want to select only men. So it depends on what your gender balance was." - E13.

6.1.4 Applicability of SP. The majority of participants were concerned about the structure of the selection rate. Eight of them said that the status quo of minority and majority would not change resulting in unfairness towards the minority (E5, E8, E9, E10, E11, 15, E16, E17). " I presume that you want equal representation and you can't do that by focusing on percentages because as you see the end of the funnel, you'll still end up with a majority and a minority." - E5. Five of them were concerned about gender, instead of merit being the focus of selection (E1, E4, E5, E7, E12). "You're not looking at the big picture of hiring the best candidates. This is statistics." - E12. "I hope that most companies won't make the decision focused solely on gender." - E5. Lastly, Four of them said that equal selection rate removes effort towards fairness or attracting more diverse applicants (E2, E10, E12, E15). "Apparently [here], we see men as higher quality than women and the risk is that you have this excuse woman. We need a woman in selection procedure. We we have more men, but we also have one woman. So we are also diverse" - E10. "I see that we have overwhelmingly male applicants or female applicants. Whatever the role is, I'm always curious why that is the gender distribution. Is it something about our job ads? Is it something about the language that we use?" - E15. Interestingly, some of them suggest that Statistical Parity is applicable only when the applicant pool is large (E5, E2) and contains only qualified

(E1, E7)and diverse (E15, E16) applicants, which would increase it's effectiveness (E2, E15). “When you create models like this, you often take the presumption that many will apply. But what will happen if you only have 3 applicants?" - E5. “If all the hard criteria is met, the percentages would make more sense and then they would feel more fair." - E1. Furthermore, many participants also suggested modifying the selection rate by making it proportional to the applicant pool representation (E5, E10, E16), making the selection rate higher for minorities (E2, E15, E16, E17), or opting for minimum number of minority candidates over a percentage (E11). “It doesn't feel fair looking at it from this perspective. I feel like the selection rate might need to be higher for women because there are fewer women if they're all qualified." - E15. “There needs to be at least one female candidate on the shortlist. It is kind of a minimum requirement of one at least one viable female candidate. Otherwise you have to keep searching. You can't only have male candidates." - E11.

Another major source of concern among the participants was the context in which the notion is applied. “It's a clear notion, but it misses lots of context." - E6. They said that applicability depends on the the composition of the existing team and type of job role, saying that a more diverse employee base and a generalist role would make Statistical Parity fair to use (E1, E5, E6, E13, E17). “I would also argue that it depends on your current team if you want to have a diverse team. If you have a team of only women, then you could argue to put more focus to men." - E6. “If you copy this model to a production facility with 150 people doing almost exact the same work, then it would be easy, really easy to implement" - E1. Some of the participants also pointed that the notion's applicability depends on the type of organization, its size and more importantly its goal (E5, E6). “You can't just replicate it towards an entire industry or even jump function or organization. It wouldn't create the effect you're looking for I think. In SMEs talent pools aren't that big." - E5.

6.2 Equal Opportunity (EO)

We obtained 88 quotations upon coding transcripts of the 17 participants. Rating distributions of the participants are shown in Figure 3.

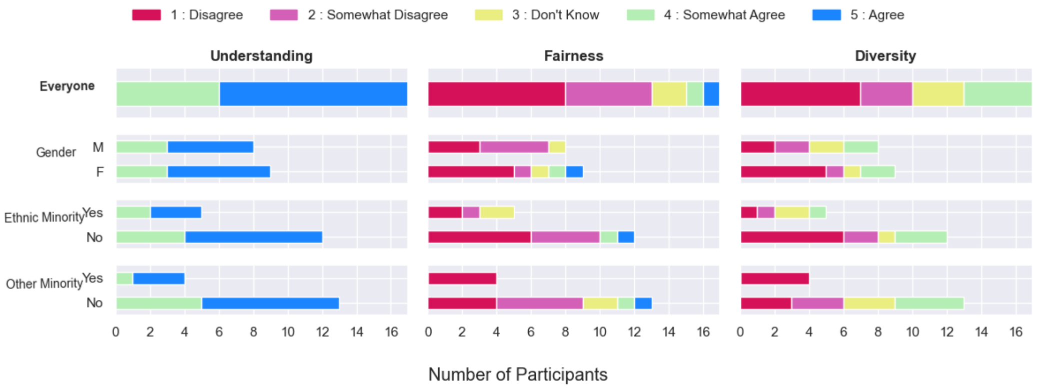


Fig. 3. **EO** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of EO, along with the distribution by *gender, ethnic minority* and *other minority* groups.

6.2.1 *Impression of EO.*

Several participants pointed out the similar underlying principle in SP and EO (E5, E9, E12), where some felt EO being less fair (E1, E2, E4, E7). “It's slightly different data, but the principle is the same." - E12. “I don't like it even more than the previous one because of this." (points to actual outcome) - E2. While all the participants understood the notion well, five participants immediately indicated their dislike or disagreement with the notion (E8, E10, E11, E12, E17). "I would never use this. It's a forced way [what] you're doing. But this is not what helps you realize non biased selection." - E12. "I don't agree. The problem is I don't agree with with selecting people on the basis of splitting up on percentages and gender. I think it will create conflict and upset people even more because it's just based upon numbers." - E11. However, a large source of dislike came from the use of actual outcome in final predictions. Nine participants either asked for reasons for the actual outcome or outright disagreed with its usage saying that it can influence a human selection-maker if the actual outcome is known (E1, E2, E4, E5, E7, E10, E13, E16, E17). "It will it will interfere with your selection method knowing what somebody else concluded." - E1. "If you ask me what happened and and why is this like this, why did they do this then? That part I don't understand." - E17.

6.2.2 Perception on fairness of using EO. A majority of 13 participants rated this notion low on fairness (score of 1 or 2) ."If you do not know the background of the decisions, I would still go for an equal selection rate of the applicant pool [SP]" - E4. While only one participant indicated the theoretical fairness of EO (E16), about seven participants expressed the notion being unfair due to the influence of actual outcome in the selections and the need more information on reasons for actual outcome (E1, E2, E4, E7, E10, E15, E17). "Has to be very clear on what basis those decisions are being made. And in this scenario, it wasn't clear. So that's not fair." - E7. "What was the selection criteria and if that's unknown to me, then this new selection doesn't seem fair because I don't have the information to make that decision" - E15. This reason also made three participants unwilling to provide a rating on fairness (E7, E14, E17). "I have no clue. I couldn't also not say if it's fair or not." - E17. "For now we are going to do a equal percentages. But when your organization needs more women or maybe more men, then you have to have other principles. But for now I cannot say anything else." - E14.

6.2.3 Perception on improving diversity when using EO. A majority of 10 participants gave low scores to the ability of EO to improve diversity. Similar patterns can be seen across genders and minorities, with no participant giving a rating of 5. From the thematic analysis of their responses, we see that many participants were certain that using EO will not improve gender balance in organizations. Six of them attributed this tothe small minority in the applicant pool (E1, E4, E7), their slim chances of being selected in the actual outcome (E4, E11) and lack of trust in the actual outcome (E1, E13, E15). "Then you most probably end up with three or four men and zero or one women." - E4. "You're copying the same bias, perhaps as the previous person" - E1.

6.2.4 Applicability of EO. The biggest source of concern for the majority of the participants was trusting the actual outcome. Six participants mentioned that bias present in the actual outcome will get copied to the prediction (E1, E2, E5, E10, E13, E16) defeating the purpose of the fairness notion. " If this (points to actual outcome) is very unfair, then it propagates unfairness." - E2. "What we do as an organization is that we never ask the opinion of the previous committee." - E10.

While 11 participants expressed concerns about using the actual outcome, three participants mentioned that they would only use the notions if the reasons behind the actual outcome is known (E6, E7, E9). "Knowing what somebody else concluded will interfere with your selection method." - E1. "The first thing that I will do is check with them. On what basis did you select those people?" - E7. "Sometimes you would have to trust that people made the right decisions and you have to move from there." - E9. One participant also said that usability would improve if there was space to disagree with the actual outcome (E16). "Instead of A1, I think I would like to interview A7 or A10 only because I'm very curious to see what they're about and just to find out if there is anything that the previous selection decisions [missed]." - E16.

The next set of concerns affecting the applicability of EO was its inability to help diversity goals. Participants said that diversity goals could not be achieved using the concept of equal selection rate. They also indicated gender, being used just for sake of diversity, thereby decreasing effort towards fairness (E9, E11, E12). "This equal selection rate for merit should be something to monitor but not to aim for." - E9. "You're confusing the situation, you're confusing things by doing this. You're just using gender to make selections without any reasoning. You can play with the percentages, you can create all kinds of different equations, but it doesn't serve justice to what you want to achieve in the end, right?" - E11. Four participants felt that just like Statistical Parity, the applicability of Equal Opportunity was affected by a skewed applicant pool, which favors the majority, suggesting a higher selection rate for minorities (E5, E7, E8, E10). "If the basis you're working from is not truly inclusive, you'll see that with every cycle that difference and imbalance magnifies." - E5. "If they're all qualified, if they can all do the job, why not select women? Because it would be very good to restore gender balance. In the long term, it's almost always better for your company." - E7. Lastly, one participant said that diversity goals can be achieved by involving multiple stakeholders, which EO currently misses (E10). "You really need these different perspectives, put them together and then you can, I think create fair principles and fair ways of working. This [EO] was probably made by one person." - E10.

6.3 Calibration (CB)

We obtained 72 quotations upon coding transcripts of the 17 participants. Rating distributions of the participants are shown in Figure 4.

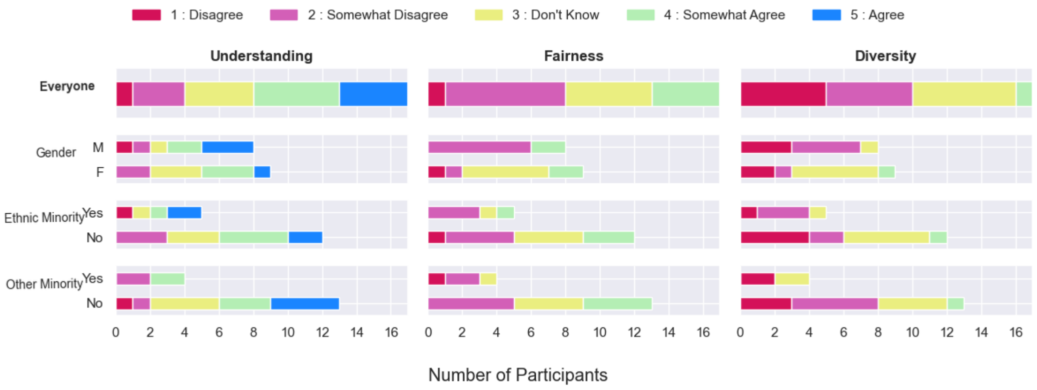


Fig. 4. **CB** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of CB, along with the distribution by *gender, ethnic minority* and *other minority* groups.

6.3.1 Impression of CB. About half of the participants rated their understanding of CB at 3 or lower. Most participants said that they found it confusing or illogical (E1, E4, E5, E6, E7, E9, E10, E13, E15, E16, E17). "I'm trying to find the logic behind this. I can can really find it" - E17. "I think this system is really confusing. I think I don't see why this would help" - E16. They said that it felt mathematical or asked clarifying questions on the scores (E1, E9). "It looks very scientific and quantitative. But I know from practice is that can be very difficult" - E9. "The points are based on what exactly?" - E17. Despite the difficulty in understanding the concept, three participants pointed out rating participants on merit is fair and can help keep human biases in check (E1, E4, E11). “If you purely look at merit based recruiting, grading candidates based on their merits is fair." - E1. Three participants extended the discussion saying they liked that applicants were compared, which can remove the myth of meritocracy (E1, E2, E17). “I would say the positive will eliminate this meritocracy myth." - E2. However, two people also disliked comparing applicants and assigning them scores (E4, E10).

6.3.2 Perception on fairness of using CB. The majority of participants rated CB as 3 or lower on fairness. However, not all participants were able to clearly express reasons behind their ratings. Two participants indicated that they needed more information and context (E13, E17). “I can't really say anything about fairness because I have no clue about the previous decisions. I really need that piece of information to. Make a statement about fairness." - E17. Three other participants felt that the process of assigning scores to applicants was unclear and they could not justify it (E4, E5, E6). “The idea of grades can help explain to people why they were selected. But in this situation, it's quite difficult because if you want to be transparent and open on it, I would say this is really difficult." - E4. “I don't understand how you can give points in this way to candidates. That's why it feels not fair if it's not transparent enough and how the points are made and how decisions are made. It lacks transparency." - E6.

6.3.3 Perception on improving diversity when using CB. The majority of the participants rated CB 3 or lower for diversity. While 10 explicitly gave low ratings of 1 or 2, 6 participants gave the ‘don't know' rating for this aspect, which may relate to this notion having been harder to grasp. Eight participants provided reasons for their ratings on diversity saying that they were either uncertain or could not see how gender balance would improve with the help of CB (E1, E2, E4, E5, E6, E9, E10, E17). “If you crystallize the process, then could it help on gender balance? Maybe. I wouldn't dare say at this point." - E5. “I think this has nothing to do with gender balance" - E10.

6.3.4 Applicability of CB. Participants expressed two major concerns regarding the applicability of CB. First, 8 participants said that the process of assigning scores was unclear making it lack transparency and logic (E2, E3, E7, E10, E11, E13, E14, E16). “It looks like it's very objective, but it's just a number and you don't know what it's based on" - E3. “That makes no sense. And I don't think it would contribute to anything. I don't think it would help gender balance. I don't think it's logical." - E11. Further, they expressed that lack of transparency made CB undesirable for use (E2, E4, E6, E11, E14). “What actually do you take into account whether something can be quantifiable, which is known to be more in favor for men like publications or grants? Then you are by default lowering the values of women." - E2. “ If I don't understand it, I can't see how it's going to help me." - E6. The second major concern came from doubts about actual outcome. Eight participants expressed theirs doubts about using it and said that it was irrelevant to use it in decisions (E2, E7, E10, E11, E12, E14, E14, E5, E7). “It can propagate bias in selection, right?" - E2. “I don't know why they made these decisions. I don't know what the selection committee was like. I have no idea. So that makes me not like using any of their previous decisions." - E15.

6.4 Fairness Through Awareness (FA)

We obtained 127 quotations upon coding transcripts of the 17 participants. The high number of quotations can be explained by half of the participants starting with FA rather than SP, and participants tending to give most feedback on the first notion. Rating distributions of the participants are shown in Figure 5.

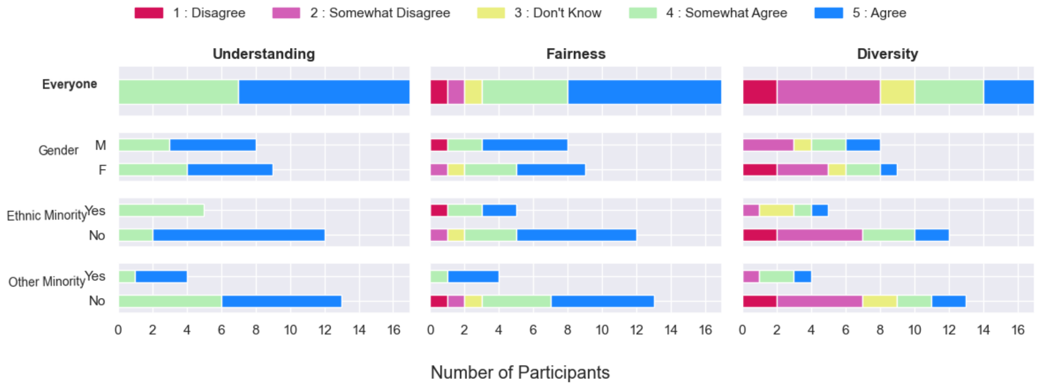


Fig. 5. **FA** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of FA, along with the distribution by *gender, ethnic minority* and *other minority* groups.

6.4.1 Impression of FA. Participants had many things to say about FA. While six participants agreed with the notion saying they were happy that both genders were treated equally (E2, E4, E5, E6, E9, E17), two participants found it difficult to comment on the notion (E1, E9). “I agree. And I would select both." - E4. “We cannot disagree with that. That is what we should aim for" - E9. On further probing 8 participants mentioned that in their experience two similar people are never the same, saying that they may differ on some aspects such as potential or soft skills (E1, E3, E4, E5, E8, E9, E13, E15). “They’re all unique, so it might differ in terms of location, match or growth in different direction" - E5. “So they had their PhD, say in 2012 and they both have, I don’t know, 16 papers. But if the woman has been on maternity leave twice, then those 16 papers means she’s been in the other time much more productive than the man. So what do you even mean with the same characteristics?" - E13.

Every participant rated their understanding at 4 or 5, indicating that they understood the definition of FA. While the majority of the participants clearly said that they understood the notion, most of them also indicated that this notion reflects the ideal end goal, cautioning its use in earlier phases of hiring (E2, E5, E8, E9, E13, E14, E16). “I understand it fully. But circumstances make it sometimes impossible to follow this principle." - E14. “I think it might work best in the beginning of the hiring process. Because, you know, in at the end of the process, you might make decisions on different data points" - E5. “I think it’s where you want to go as an end goal but to drive the change for unconscious bias, need to create more opportunities for women" - E8.

6.4.2 Perception on fairness of using FA. Majority of the participants rated fairness of FA at 4 or higher. Interestingly, they reflected on many dimensions to assess the fairness of FA. While four participants referred to FA as being principally fair (E6, E8, E10, E11), several others provided reasons for calling the notion unfair. “The principle in itself is rationally fair. The world that we live in is not." - E10. Participants called FA unfair because it is not equitable, cannot achieve diversity goals, disadvantage to minorities (E2, E7, E8, E15). “I think given the gender imbalance for which we want to correct, then I don’t think this is fair " - E2. “This is a utopia idea. I think it’s still inequitable because the female will always have a bias, that she is considered less fit for the job if everything else is the same" - E8. Further, participants reflected that while FA might be fair towards applicants (E1, E3, E5, E7, E12, E16, E17), minorities groups might still find it unfair (E1, E14, E15) and they felt that it was more important to fix historical justice because majorities won’t be severely affected if minorities are given more opportunities (E2, E8, E11, E15, E17). “It’s difficult for the male candidate, but we have a kind of historical injustice that has to be fixed. It doesn’t mean we only select female applicants, right?" - E11. Lastly, four participants found it quite difficult to rate fairness of FA saying that fairness depends on many factors such (E9, E10, E11, E13). “I find it very hard to answer because I need a more clear definition of what the the same characteristics are." - E13. “It’s fair within the scope of the context, which I would define but as a general statement, it’s not fair." - E11.

6.4.3 Perception on improving diversity when using FA. In contrast to ratings on fairness, ratings on diversity for FA are divided. While three participants said that FA could help organizations improve gender balance (E4, E5, E15), seven other participants said that FA could not improve gender balance because it is prone to cultural cloning, skewed applicant pools and minorities self-selecting themselves for job applications (E2, E6, E8, E9, E10, E14, E16). “If you have a black woman and you have a white man and the previous 20 people that did the job were all white men, then the white man will be hired again. This is how our brain works." - E10. “If you want to improve gender balance in an organization, then sometimes you cannot have this fair criterion. Sometimes you have to let that go." - E9. Lastly, five participants indicated that the notion could help in some contexts, while being supported by other measures within the organization (E3, E4, E7, E11, E17). “Yes, it will help improvement, but it’s not the only thing." - E11. “It depends on so many things. I don’t know how many women or men [sic] employees the organization has." - E17.

6.4.4 Applicability of FA. The majority of the participants said that FA is quite theoretical and difficult to apply in practice (E1, E2, E5, E8, E9, E10, E11, E13). “It’s the most fair thing to do, but it doesn’t make any sense" - E1. “In the perfect world where everybody would be treated equally and we have equal opportunities, then this principle is great but we are not in that world right now." - E10. “On paper, it may look very easy, but in practice it’s not." - E9. Additionally, five participants also indicated that the notion benefits the majorities more than it benefits the minorities and cannot select applicants who have different characteristics than previous employees (E1, E3, E4, E8, E10). “It could be that a male applicant has more profit of this fairness notion than a woman because maybe a woman was pregnant a couple of times" - E3. “So if you have always had white men of a certain age with a certain background, certain studies, and they were always doing the job in a good way, you will pick a person that is the same as all the people that did it before" - E10. Lastly, some participants mentioned that the notion would be applicable only in organizations that are diverse (E7, E10, E11, E14). “When [your organization] is balanced and you are in that sense equal, you’re also giving equal opportunities to everybody. Then you can apply this principle of fairness for sure." - E10.

6.5 Counterfactual Fairness (CF)

We obtained 72 quotations upon coding transcripts of the 17 participants. Rating distributions of the participants are shown in Figure 6.

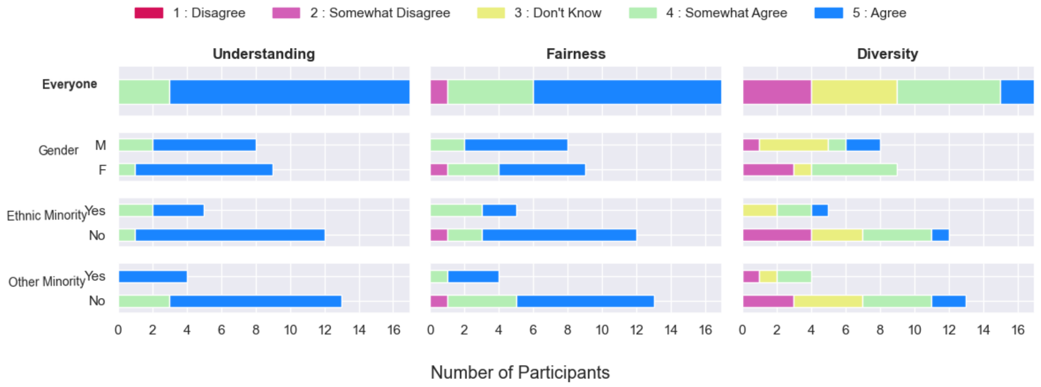


Fig. 6. **CF** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of CF, along with the distribution by *gender, ethnic minority* and *other minority* groups.

6.5.1 Impression of CF. CF was very well understood, with a majority of the participants rating understability at 5. There also was high agreement with the notion (E1, E3, E6, E7, E8, E9, E10, E12, E13, E14), with two participants expressing curiosity about trying out the notion (E1, E16). “I agree that changing only gender should not affect the selection decision." - E14. “I really like this especially if you test it." - E1. Additionally, some participants mentioned that such a notion can help keep discrimination in check (E1, E2, E16). “You’re not pushing more women forwards or more men forwards. You’re just checking bias" - E1. Lastly, few participants indicated the similarity of CF and FA (E1, E7) “I think there’s [sic] they’re ethically speaking the same, right?" - E7.

6.5.2 Perception on fairness of using CF. Perception of fairness and diversity of using CF elicited few responses from the participants. Five participants said that the notion is fair in principle but is not very practical or useful (E5, E10, E10, E14, E17). “The principle in itself is very fair. But putting it in the context of our world becomes very complex." - E10. Another set of four participants felt the notion is fair because it is purely based on merit and could help improve the selection of minorities (E7, E9, E11, E15). “I would rate this as fair because nothing has changed about the qualifications of the candidate at this point." - E15.

6.5.3 Perception on improving diversity of using CF. Ratings are divided for diversity across the Likert scale and there is no consensus on it. Majority of the participants were unsure whether CF would help improve the gender balance in organizations and indicated the using other measures instead (E2, E3, E4, E5, E6, E8, E9, E10, E12, E14). “Can it help improve gender balance? I’m not sure. This notion, not by itself, no." - E6. “No, it won’t help. When you want to have more women in your organization, Then sometimes you have to [use other approaches]" - E14. “I don’t think in in terms of fairness or diversity it can be improved. At least temporarily, we might need different measures." - E5. Few participants indicated that depending on the context, it could help diversity (E11, E16). “We want to get more males for secretarial support. There, it would contribute to helping the gender balance." - E11.

6.5.4 Applicability of CF. The biggest concern about CF expressed by participants was related to gender. Four participants said that they were unsure about gender or another sensitive feature being part of the selection process when CF is used (E1, E6, E9, E17). “I think it’s all information that we do not need for open, transparent and merit based recruiting. It might help with diversity, though. And sometimes a big age gap can be seen as something less positive too" - E1. While, three other participants said that diversity is a quality and such features can be helpful in selection (E3, E4, E10). "To me and to our organization, diversity is also a quality. So, if the norm in your organization is female and the applicant is male, then that is also a quality of the person" - E10. "Men and women bring different perspectives which you should also consider. You don't see this if you remove the gender." - E3. Two additional participants indicated low applicability of CF, if diversity is a goal of the company (E7, E11).

On the applicability of CF, four participants said that improvement would be quite slow because the notion favors the majority (E2, E5, E13, E15). "You have marginalized groups that are like 10 - 0 behind. So you can use the notion, but that would mean that we would have to go to many cycles to reach to reach true fairness levels. So it's not always applicable." - E5. Lastly, two participants indicated that the notion is not bidirectional, meaning that it was not logical to show majority as a minority applicant because they would still be favored (E2, E13). "I think what occurs more is that you would have females who are not selected and then if they were presented as a male candidates, they would be selected and not the other way round." - E2. "You can never have the same candidate's gender flipped because a man cannot become pregnant, so you cannot have a gender neutral CV because how do you account for pregnancy leave then." - E13.

6.6 Changes in the ratings

Having discussed and rated all the notions, participants were asked to rate the same notions on Fairness (F) again. This second round of ratings was obtained to mitigate order effects, while at the same time nudging participants to now comparatively consider the different notions they saw passing by. In re-rating, participants could not revisit their initial ratings. Figure 7 visualizes to what extent initial and final ratings by participants changed for each fairness notion.

Comparatively, SP sees many shifts in ratings, possibly due to this having been one of the starter notions. Participants who initially interpreted SP as unfair, tend to still feel the same or move to a more positive rating in the second round. At the same time, several participants who used to find SP somewhat fair are more negative in retrospect. Overall, participants are divided in their final judgments, showing the most uniform distribution over the 5 ratings out of all notions.

Where EO initially invoked largely negative ratings, participants becomes milder towards this notion at the end. Still, a majority of ratings remain on the negative side of the Likert scale.

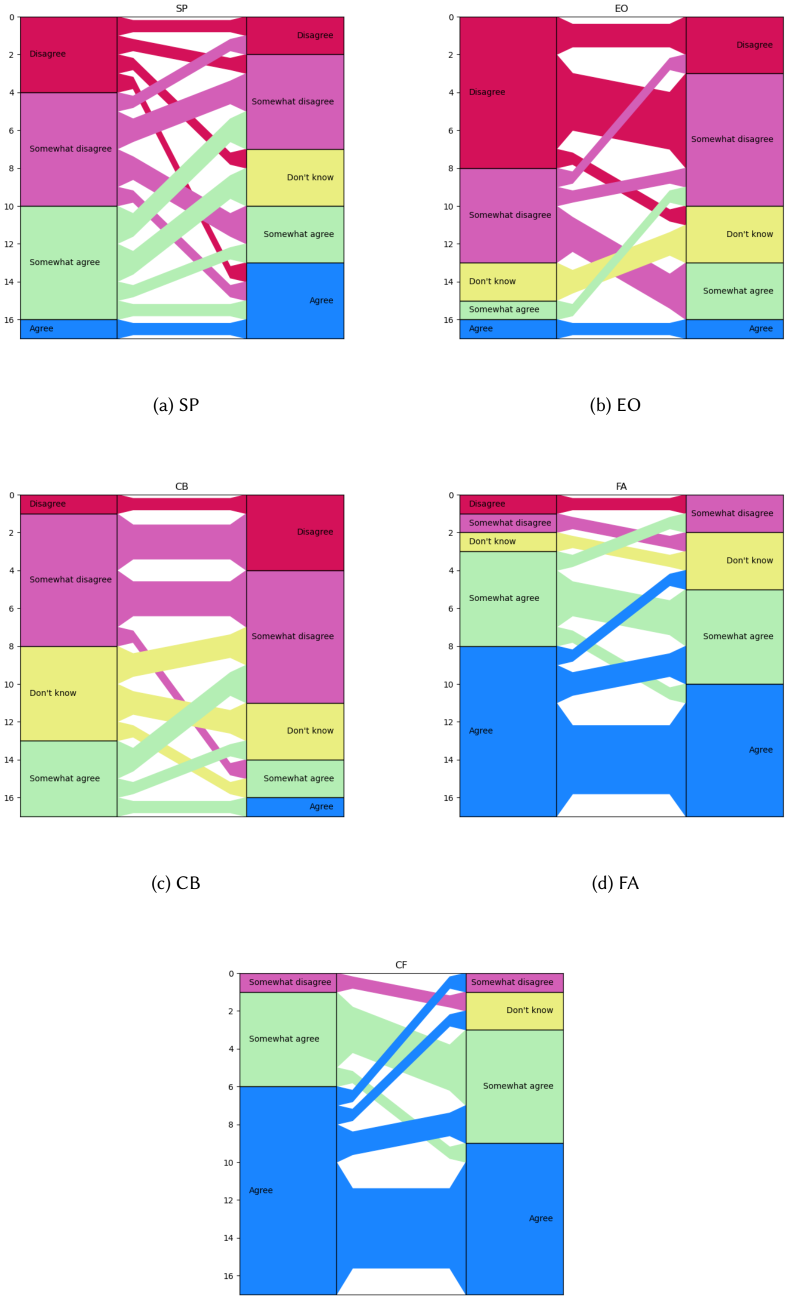
Where CB was hard to grasp, and several participants initially were hesitant to rate it for fairness, in the end, participants more explicitly take a stance, mostly skewing towards the negative side of the Likert scale.

The initial response to FA with regard to fairness was more positive than in the end. Still, this notion overall retains a majority on the positive side of the Likert scale, and the most negative rating disappears.

As with FA, participants initially were more positive about the fairness of CF than at the end, although the majority remains on the positive side of the Likert scale.

7 DISCUSSION

In the previous section, we gave a rich overview of our participants' responses relating to our research questions for each of the chosen mathematical fairness notions. Zooming out, our observations lead to a few more insights.



(a) SP

(b) EO

(c) CB

(d) FA

(e) CF

Fig. 7. Alluvial plot showing changes in the ratings for all fairness notions. Initial ratings on the left move to final ratings on the right.

7.1 Considerations of Understanding, Fairness, Improving Diversity and Applicability

7.1.1 Some notions are more complicated to grasp. Where overall, participants had high understanding of the fairness notion explanations, CB (that already was found harder to grasp during our design phase) remained more difficult to understand. The lower degree of understanding also led to more hesitant responses regarding fairness, diversity and applicability. This aligns to the findings in [39], where lay people found it difficult to judge fairness of complex notions, and ended up choosing the simplest notion, statistical parity, as the most fair option. In our case, the latter however did not happen. This may have to do with domain experts being concerned about SP conflicting with considerations of merit, which are important in hiring, where [39] considered other application domains (recidivism prediction and skin cancer diagnosis), while the raters were no domain experts. At the same time, in our case, SP did not stand out as obviously simpler to understand in comparison to other notions.

7.1.2 More context may be needed to judge fairness. Most participants found it quite difficult to provide a rating for fairness. They said that that they found the question too difficult, as they did not have a clear set definition for fairness themselves . “Can you call a process that focuses on education and experience in itself fair? I think that’s a difficult question to answer" - E5. “I find it quite difficult answering your questions. I mean, I would like to separate the fairness and the judging" - E13. “It’s quite difficult to say if it’s fair or not because you think the answer is yes, but I don’t think it is". - E16. Most participants provided their rating after talking about multiple dimensions of fairness, asking whether it was fair to organizations or applicants. “It depends how you define fairness. Yeah. It’s depends what you understand on the word fairness. Okay. So I’d say don’t know because I don’t know what you mean by fairness, because fairness is a, is a subjective term. So objectively is not fair. Subjectively it’s fair. Depends how you want to view it." - E11. This points to contextualizing and defining the scope of fairness in future studies, as people find it quite subjective. This is aligned with findings from [22], where participants had no consensus on fairness and their discussion went beyond discrimination.

7.1.3 Minoritized groups may bring specific, different stances on how to improve diversity. In [42], it is shown how algorithmic fairness evokes negative emotions from minority groups regarding racial and economic justice. We did not explicitly frame our questions as questions on algorithmic fairness to our participants, and our sample has been too small to warrant quantitative conclusions. However, in the reactions to our question on whether given fairness notions would help in improving diversity, we did seem to see different, stronger responses from those identifying with minority groups. Participants sometimes refused to answer this question, saying that they were forced to rate an option they did not agree with. “So it’s it’s [sic] a bit of a dilemma. There’s no there’s no clear answer on this. So you’re trying to force me into a box and you can’t." - E11. “I’m missing quite a bit of context. It’s really hard. I can’t. I can’t give a proper answer. I find it really hard. These kind of notions, the way we are thinking about this diversity and inclusion. You need to have these numbers, right? Because then you have a better understanding. But for me it’s too simple, too much simplified actually" - E6. In our study, most of the professionals we reached did not identify as belonging to a minority. In the future, it will be interesting to recruit a more balanced sample on this, and investigate to what extent there indeed are response differences between those identifying as minoritized and non-minoritized. To further strengthen a case for this, for the sensitive attribute on which he had a reasonably balanced sample (gender), on closer examination, we see that proportionally more women compared to men were skeptical about all the notions’ abilities to improve diversity.

7.1.4 Fairness, Improving Diversity and Applicability do not necessarily go hand in hand. For individual fairness notions, many participants took time to suggest that while the notion was fair, it would not help diversity goals of their organizations, while some were even surprised. The tension between fairness and diversity is reflected in the high ratings for fairness but low ratings for diversity. "If you want to improve gender balance in an organization, then sometimes you cannot have this fair criterion. Sometimes you have to let that go." - E9. Similarly, as already discussed in the reactions to FA and CF, what is considered fair may not be applicable or defensible in practice.

7.2 Domain-specific considerations

7.2.1 Discomfort on selection based on sensitive attributes. Some important tensions seen throughout the discussion with participants was whether to include a sensitive feature such as gender so explicitly in the selection process. "I don't think it's fair to select on gender." - E17. "I think that's not really fair. When you when it's, for instance, the first thing you look at when you are starting the hiring procedure." - E4. Another concern that emerged was the fear of selecting unqualified female candidates for the 'sake' of fairness. "It feels a lot like the woman is there as an excuse." - E10. "You should always select based on merit and merit doesn't come presented in gender." - E1. While most participants deliberated both the concerns on gender and merit, many women explicitly expressed that general notion of merit favored men and more inclusive takes on merits are needed before such notions could be applicable and called fair. "We see also that women who, if they have children in their careers, that it has an impact on maybe a number of publications or whatever, and that should not hinder them in making progress or being hired because male applicant has maybe more publications". - E8. "I find this topic is just very complex because when you look at experience, we see that because systematically, women get less chances in work than men do, the women also have less experience and therefore, have less strong resumes. But that is a system and you want to break through that system because we always say fix the system, not the women". - E2.

7.2.2 Domain-specific context and critical nuances are missing in a mathematical fairness notion. Many participants expressed that fairness notions miss many critical nuances and context to questions of early selection. Not having access to these nuances and context, such as team composition, type of industry, goals, etc., hinders them from being certain of their answers, and makes them more resistant to see fairness notions as potential ways to get more explicit and standardized policy. "The most important thing is always context. Because you can't say that something is fair without having something to bounce it off of. Like the company or the status quo of the company, the hiring process itself. There are many more factors than just equality, than just minority majority or male female. It should be more dynamic and not like a fixed way of approaching [fairness]" - E6. "It needs more clarity. So if you say within the context of the needs of [diversity], it's fair. If you would say in context of the male candidate, it would probably be unfair." - E11. More effort will need to be spent on connecting these notions further with the domain. Here, it is likely that a 'one size fits all' approach will not exist.

7.3 Further considerations

7.3.1 Ratings do not reflect sentiment. Even if a larger sample of participants and ratings would be reached, it is important to not only look at numerical ratings, but also at the rationale behind them. Most of the participants felt that they could not disagree with individual fairness notions of FA and CF and rated it as fair. However, their language suggested some restrain or skepticism. "The principal is fair, but the human brain is not fair.So yes, I'm sorry I find is a very hard question. I suppose the principle is fair." - E10. Sentiments of disinterest in continuing the discussion with all the notions, despite being enthusiastic on talking about such topics, reflects that the ratings do not convey the full picture. “if you start doing this, is it fair? But the notion and mindset of fairness needs to change. You need to first understand who is the minority, do they actually get the same opportunity? No." - E8. “in the perfect world where everybody would be treated equally and we have equal opportunities, then this principle I think is great but we are not in that world right now." - E10.

7.3.2 Participants are intrinsically motivated to engage with questions of fairness. Our interviews demanded considerable time investment from our participants (60-75 minutes), sometimes requiring for participants to return for a second sitting to complete the session. At the same time, no resources were available to compensate participants for their time. Still, participants were eager and intrinsically motivated to participate, return, and stay in the loop on the authors’ findings. This has been a promising observation, suggesting that further collaboration between researchers and practitioners would be a logical step to take.

7.4 Limitations and Future Work

Several limitations can be identified in our current work. First of all, while our non-technical translations of mathematical fairness notions were iteratively developed with feedback from independent colleagues, they still can further be refined, and more thoroughly be evaluated on indeed being an understandable and sufficiently accessible depiction of the fairness notions of interest.

Furthermore, many more fairness notions have been proposed in literature, and we only studied a few of them. More notions may need to be investigated, while at the same time, it is likely that all of the notions may be too rigid, too inflexible and too distanced from application contexts to be considered as a sufficiently acceptable explicit reference for implementing diversity policy. It is imaginable that different notions need to be combined. Generally spoken, more explicit discussions also need to be held on the degree to which potential candidate rankings in current selection processes can be trusted not to carry undesired biases, and whether all relevant facets to a candidate being qualified already would be sufficiently captured.

Continuing on discussions on multifacetedness, our current work also only considered one sensitive attribute: gender. Even while in practice, it seems very hard to monitor for other sensitive attributes, it will be worthwhile to investigate considerations on other sensitive attributes, and furthermore consider aspects of intersectionality, which may lead to extra adverse effects on those belonging to multiple minoritized groups [13].

Generally, the question of fairness needs to be considered in the context of its application. While we zoomed in to early candidate selection in hiring, even within this scope, refinement and contextualization is needed, which also may include more explicit connection to surrounding elements and stakeholders in the pipeline. To make the discussion more tangible and recognizable, it will be worthwhile to not only discuss theoretical, fictional examples, but integrate these more strongly with cases and infrastructures from actual practice. However, in doing this, the technological intervention still would need to be considered next to non-technological organizational aspects and facilities, such as company reputation, awareness trainings, and the facilitation of inclusive work environments.

Finally, choices of fairness notions reflect choices of world views, and challenges of untangling and balancing these. With regard to this, academics in the field of Ethics have been more explicitly articulate and aware about this. In our current work, we did not integrate this awareness yet. However, in future refinements of non-technical translations of fairness notions, it will be useful to more consciously take this into account, and more explicitly include this in the presentation of the notions.

Presently, through our study, we have given insights in participant responses to our translations of different fairness notions. While this describes how participants currently think about the different notions, we did not yet take a stance on consequent preferred notions to adopt or integrate. This also has to do with the observed tensions between fairness, diversity improvement and applicability. In addition, with our participants having been confronted with these types of notions for the first time, they may not oversee the potential impact of the different notions yet. For example, it is striking that the individual fairness notions FA and CF are rated less negatively than the group fairness notions on their capability of improving diversity. However, in credit risk scoring, it actually has been shown that implementing individual fairness notions may amplify inequality between groups that already are far apart, thus being particularly disadvantageous for members of the disadvantaged group [6].

Where hiring is a different application domain from credit risk scoring, being rejected in a job selection process could be seen as being denied an opportunity for financial stability. Similarly, linking back to ethics, that what respondents currently prefer may not actually be the preferred course of action is reminiscent of the distinction between social acceptance and ethical acceptability. Arguments that these both need to be considered when policy is to be set [40] can easily be extended to questions of fairness in data-driven decision-making. Similarly, in the disagreements seen between our participants, and hesitance of participants to rate when context is missing, we see aspects of conceptual and epistemic normative uncertainties [41].

When starting the research leading to this paper, initially, the authors had intended to work on (quantitative) fairness monitoring tooling for early selection stages. However, as our investigations show, deeper qualitative understanding of the problem space is key before any quantitative tool will have relevance. Considering the responses of our participants, even if good data would be available, it would not have made sense to implement functionality from common fairness toolkits at this point in time. In finding ways to understand where data-driven tooling may be helpful—or even before that, where the room is to concretize towards more transparent and actionable improved selection policies—as we pointed out, more connections between expertise in different academic domains can and should be investigated, in active collaboration with practitioners, designers and policy-makers.

REFERENCES

[1] Ifeoma Ajunwa. 2019. An auditing imperative for automated hiring.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications, 254–264.

[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104, 671.

[4] Rachel K. E. Bellamy et al. 2018. AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv:1810.01943 [cs]*, (Oct. 3, 2018). Retrieved Mar. 29, 2022 from http://arxiv.org/abs/1810.01943 arXiv: 1810.01943.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv:1706.02409 [cs, stat]*, (June 7, 2017). Retrieved Jan. 21, 2022 from http://arxiv.org/abs/1706.02409 arXiv: 1706.02409.

[6] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 514–524.

[7] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. preprint. SocArXiv, (Jan. 31, 2018). DOI: 10.31235/osf.io/9wqxr.

[8] Sarah Bird, Miro Dudık, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Tech. rep. MSR-TR-2020-32. Microsoft, (May 2020). https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[9] J Stewart Black and Patrick van Esch. 2020. Ai-enabled recruiting: what is it and how should a manager use it? *Business Horizons*, 63, 2, 215–226.

[10] Miranda Bogen and Aaron Rieke. 2018. Help wanted: an examination of hiring algorithms, equity, and bias.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 2, (Jan. 2006), 77–101. DOI: 10.1191/1478088706qp063oa.

[12] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: a qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19: CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk, (May 2, 2019), 1–12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300271.

[13] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[14] Tomas Chamorro-Premuzic and Reece Akhtar. 2017. Should companies use ai to assess job candidates? (2017). https://hbr.org/2019/05/should-companies-use-ai-to-assess-job-candidates.

[15] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *arXiv:1610.07524 [cs, stat]*, (Oct. 24, 2016). Retrieved Jan. 24, 2022 from http://arxiv.org/abs/1610.07524 arXiv: 1610.07524.

[16] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv:1810.08810 [cs, stat]*, (Oct. 20, 2018). Retrieved Jan. 6, 2022 from http://arxiv.org/abs/1810.08810 arXiv: 1810.08810.

[17] Bo Cowgill. 2019. Bias and productivity in humans and machines. *Columbia Business School Research Paper Forthcoming*.

[18] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19: 24th International Conference on Intelligent User Interfaces. ACM, Marina del Ray California, (Mar. 17, 2019), 275–285. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302310.

[19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness through awareness. *arXiv:1104.3913 [cs]*, (Nov. 28, 2011). Retrieved Jan. 24, 2022 from http://arxiv.org/abs/1104.3913 arXiv: 1104.3913.

[20] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *arXiv:1609.07236 [cs, stat]*, (Sept. 23, 2016). Retrieved Mar. 10, 2022 from http://arxiv.org/abs/1609.07236 arXiv: 1609.07236.

[21] Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics*, 148–170.

[22] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: a case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. the 2018 World Wide Web Conference. ACM Press, Lyon, France, 903–912. ISBN: 978-1-4503-5639-8. DOI: 10.1145/3178876.3186138.

[23] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv:1610.02413 [cs]*, (Oct. 7, 2016). Retrieved Dec. 8, 2021 from http://arxiv.org/abs/1610.02413 arXiv: 1610.02413.

[24] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, Virtual Event Canada, (Mar. 3, 2021), 375–385. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445901.

[25] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in learning: classic and contextual bandits. *arXiv:1605.07139 [cs, stat]*, (Nov. 7, 2016). Retrieved Mar. 7, 2022 from http://arxiv.org/abs/1605.07139 arXiv: 1605.07139.

[26] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. [n. d.] Preventing fairness gerrymandering:auditing and learning for subgroup fairness, 9.

[27] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics*, 133, 1, 237–293.

[28] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807 [cs, stat]*, (Nov. 17, 2016). Retrieved Jan. 18, 2022 from http://arxiv.org/abs/1609.05807 arXiv: 1609.05807.

[29] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

[30] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5, 1, (Jan. 1, 2018), 2053951718756684. Publisher: SAGE Publications Ltd. DOI: 10.1177/2053951718756684.

[31] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A human-centered approach to algorithmic services: considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17: CHI Conference on Human Factors in Computing Systems. ACM, Denver Colorado USA, (May 2, 2017), 3365–3376. ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025884.

[32] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. 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*, 166–176.

[33] Cynthia C. S. Liem, Markus Langer, Andrew Demetriou, Annemarie M. F. Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph. Born, and Cornelis J. König. 2018. Psychology meets machine learning: interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. The Springer Series on Challenges in Machine Learning. Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel A. J. van Gerven, (Eds.) Springer, 197–253. ISBN: 978-3-319-98130-7. DOI: 10.1007/978-3-319-98131-4_9.

[34] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. On the applicability of ML fairness notions. *arXiv:2006.16745 [cs, stat]*, (Oct. 19, 2020). Retrieved Nov. 3, 2021 from http://arxiv.org/abs/2006.16745 arXiv: 2006.16745.

[35] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Prediction-based decisions and fairness: a catalogue of choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, 1, (Mar. 7, 2021), 141–163. arXiv: 1811.07867. DOI: 10.1146/annurev-statistics-042720-125902.

[36] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. Conference on Fairness, Accountability, and Transparency. (2018). https://www.youtube.com/watch?v=jIXIuYdnyyk.

[37] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20: Conference on Fairness, Accountability, and Transparency. ACM, Barcelona Spain, (Jan. 27, 2020), 469–481. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372828.

[38] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. 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*. AIES '19: AAAI/ACM Conference on AI, Ethics, and Society. ACM, Honolulu HI USA, (Jan. 27, 2019), 99–106. ISBN: 978-1-4503-6324-2. DOI: 10.1145/3306618.3314248.

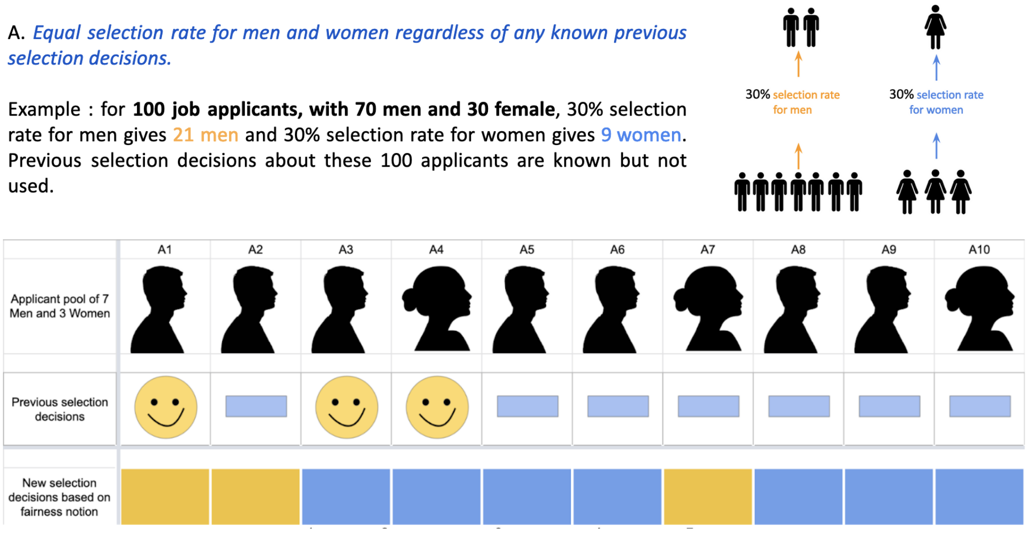
[39] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: a descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, Anchorage AK USA, (July 25, 2019), 2459–2468. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330664.

[40] Behnam Taebi. 2016. Bridging the gap between social acceptance and ethical acceptability. English. *Risk Analysis: an international journal*. DOI: 10.1111/risa.12734.

[41] Behnam Taebi, Jan H. Kwakkel, and Céline Kermisch. 2020. Governing climate risks in the face of normative uncertainties. English. *Wiley Interdisciplinary Reviews: Climate Change (Online)*, 11, 5. DOI: 10.1002/wcc.666.

[42] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18: CHI Conference on Human Factors in Computing Systems. ACM, Montreal QC Canada, (Apr. 21, 2018), 1–14. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174230.

A FINAL DESIGNS

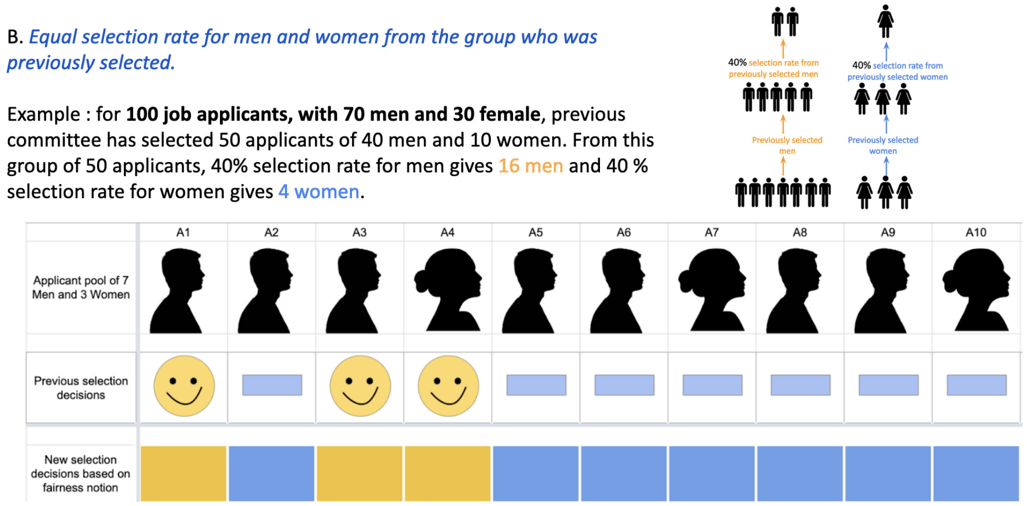


A. *Equal selection rate for men and women regardless of any known previous selection decisions.*

Example : for **100 job applicants, with 70 men and 30 female**, 30% selection rate for men gives 21 men and 30% selection rate for women gives 9 women. Previous selection decisions about these 100 applicants are known but not used.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Applicant pool of 7 Men and 3 Women	👨	👨	👨	👩	👨	👨	👩	👨	👨	👩
Previous selection decisions	☺	▭	☺	☺	▭	▭	▭	▭	▭	▭
New selection decisions based on fairness notion	🟨	🟨	🟦	🟦	🟦	🟦	🟨	🟦	🟦	🟦

(a) Final design of the toy example for SP



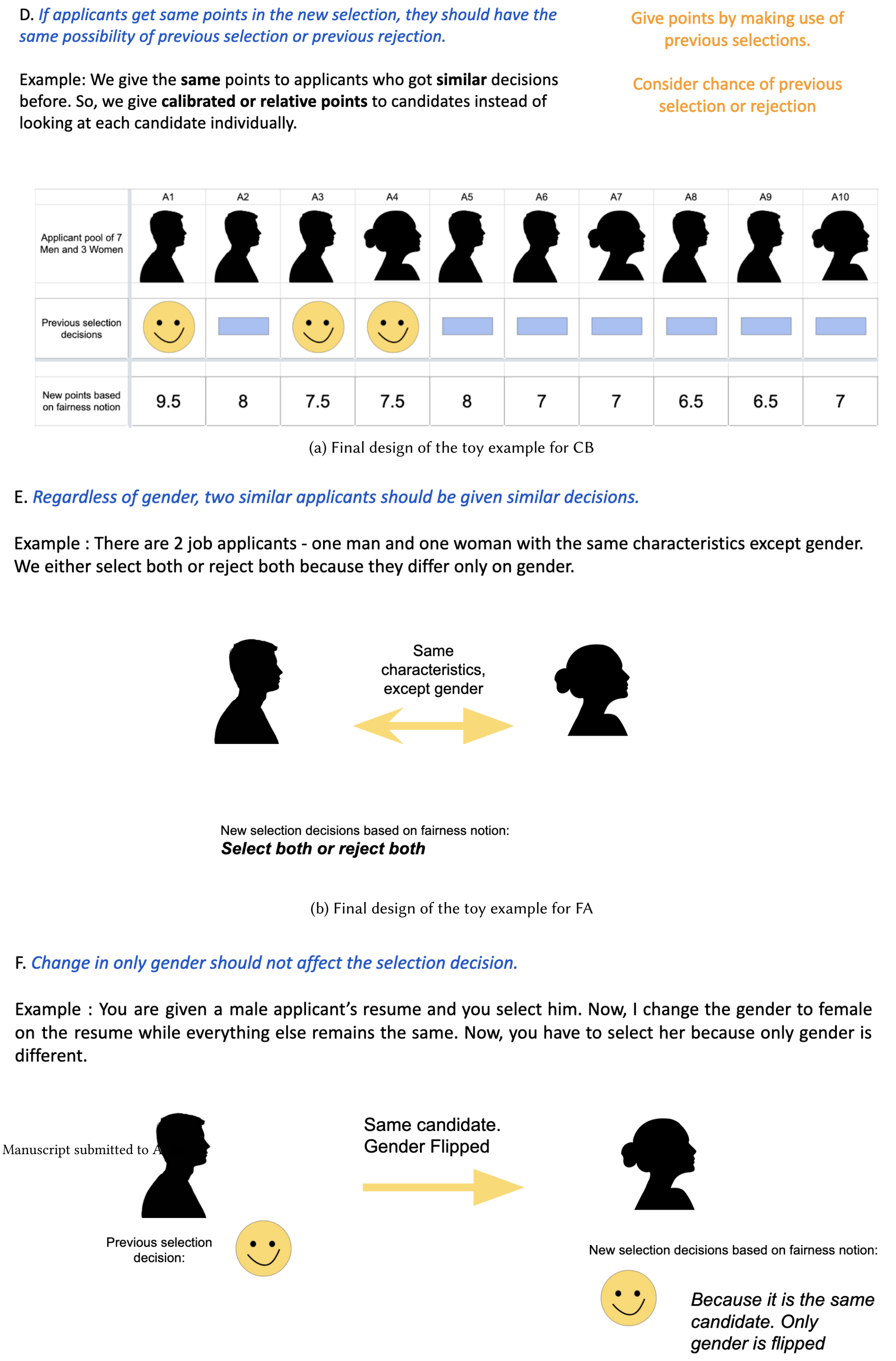
B. *Equal selection rate for men and women from the group who was previously selected.*

Example : for **100 job applicants, with 70 men and 30 female**, previous committee has selected 50 applicants of 40 men and 10 women. From this group of 50 applicants, 40% selection rate for men gives 16 men and 40 % selection rate for women gives 4 women.

40% selection rate from previously selected men
40% selection rate from previously selected women
Previously selected men
Previously selected women

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Applicant pool of 7 Men and 3 Women	👨	👨	👨	👩	👨	👨	👩	👨	👨	👩
Previous selection decisions	☺	▭	☺	☺	▭	▭	▭	▭	▭	▭
New selection decisions based on fairness notion	🟨	🟦	🟨	🟨	🟦	🟦	🟦	🟦	🟦	🟦

(b) Final design of the toy example for EO



D. *If applicants get same points in the new selection, they should have the same possibility of previous selection or previous rejection.*

Example: We give the **same** points to applicants who got **similar** decisions before. So, we give **calibrated or relative points** to candidates instead of looking at each candidate individually.

Give points by making use of previous selections.

Consider chance of previous selection or rejection

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Applicant pool of 7 Men and 3 Women										
Previous selection decisions	🙂	▬	🙂	🙂	▬	▬	▬	▬	▬	▬
New points based on fairness notion	9.5	8	7.5	7.5	8	7	7	6.5	6.5	7

(a) Final design of the toy example for CB

E. *Regardless of gender, two similar applicants should be given similar decisions.*

Example : There are 2 job applicants - one man and one woman with the same characteristics except gender. We either select both or reject both because they differ only on gender.

New selection decisions based on fairness notion:
Select both or reject both

(b) Final design of the toy example for FA

F. *Change in only gender should not affect the selection decision.*

Example : You are given a male applicant's resume and you select him. Now, I change the gender to female on the resume while everything else remains the same. Now, you have to select her because only gender is different.

New selection decisions based on fairness notion: 🙂

Because it is the same candidate. Only gender is flipped

(c) Final design of the toy example for CF