“It's the most fair thing to do, but it doesn't make any sense”
Perceptions of Mathematical Fairness Notions by Hiring Professionals

Sarkar, Priya; Liem, Cynthia C.S.

DOI
10.1145/3637360

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the ACM on Human-Computer Interaction

"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, Delft University of Technology, The Netherlands
CYNTHIA C. S. LIEM, Delft University of Technology, The Netherlands

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. In this, 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, personnel selection

ACM Reference Format:
Priya Sarkar and Cynthia C. S. Liem. 2024. "It's the most fair thing to do, but it doesn't make any sense": Perceptions of Mathematical Fairness Notions by Hiring Professionals . *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 83 (April 2024), 35 pages. https://doi.org/10.1145/3637360

1 INTRODUCTION

In present-day society, many aspects of unfairness and inequality exist. For example, in credit risk scoring, considerable disparities exist between risk score distributions of different racial groups, which will affect the economic opportunities of members from these groups [42]. In academic promotions, the career advancement of women may be hampered by gendered stereotypes, without these stereotypes being recognized and acknowledged [97]. In hiring, discrimination towards ethnic minorities has been observed in different countries [43], where reasons both lie with different ways of resume presentation, as well as human capital disadvantages.

Concerns about unfairness in human resources processes (e.g., workplace circumstances, hiring policies, career advancement opportunities) are increasingly recognized as an organizational operational concern. In response, Diversity, Equity and Inclusion (DEI) frameworks and offices are increasingly established, that seek to implement fairness-promoting strategies in the policies of organizations [28, 38, 20]. The people working on implementation of such policies tend to need to

Authors' addresses: Priya Sarkar, priyasarkar.contact@gmail.com, Delft University of Technology, Postbus 5, Delft, The Netherlands, 2600 AA; Cynthia C. S. Liem, c.c.s.liem@tudelft.nl, Delft University of Technology, Postbus 5, Delft, The Netherlands, 2600 AA.

ACM 2573-0142/2024/4-ART83
https://doi.org/10.1145/3637360

push against an established and institutionalized status quo, with role ambiguity, lack of support, and tokenism (being singled out as a 'token' representing a minority, because of being a member of this minority) [107]. This makes them vulnerable to burnout [84, 104], while at the same time, pressure exists to make DEI interventions more evidence-based, while little data is still available on their impact [104].

In parallel, society has turned increasingly digital, and algorithmic systems have increasingly been proposed as ways to automate decision-making and prioritization processes. Often, these are based on machine learning, and optimized for recognizing and reproducing statistical patterns in existing data. Various minoritized authors have warned that these technologies have predominantly been designed by privileged white males, who have been oblivious to large-scale deployment having adverse consequences for vulnerable populations [79], with sexist [26] and racist [78, 6] impacts.

Within the machine learning (ML) community, unfair social impact of predictive methods has also increasingly been recognized as a concern. In response, on the algorithmic side, research on ML fairness gained traction, which led to a broad range of mathematical fairness notion proposals (e.g., [23, 42, 8, 55, 21, 37, 30, 57, 52]), that can quantitatively be used as optimization or auditing criterion in data-driven systems. To ease adoption, several toolkits implementing these notions have been proposed, such as the IBM AI Fairness [5] and FairLearn [11] 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 [87]. However, the problem is more complicated, as translating real-world problems to data and problem framing compatible with machine learning frameworks is a highly non-trivial matter [81, 27, 44, 82].

This raises questions of operationalization: how do design requirements concretely translate into technical choices, and which fairness notion would one choose to implement? In applications of algorithmic decision-making such as in credit risk scoring [42, 90, 89], criminal recidivism prediction [2, 34], or hiring [98, 13, 102], operationalization choices intended to increase fairness were shown to actually perpetuate inequality, adversely affecting minority populations. The various possible mathematical fairness notions capture fundamentally different world views [74, 36], that mathematically cannot be satisfied at the same time [36, 77]. Furthermore, questions of measurement bias and overarching questions of experimental validity have not trivially been included in machine learning methodology [47, 64, 44, 27, 100].

It also should be noted that the algorithmic take on fairness has been led by computer and data scientists. However, with questions of fairness and (in)justice being situated in more systemic social scenarios, it will be appropriate to rather contextualize algorithmic fairness discussions at interdisciplinary crossroads [27, 91, 60, 71]. 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 [64, 95].

In this article, we seek to initiate such an interdisciplinary discussion, by exploring the alignment of organizational representatives involved in hiring and DEI processes on five different, commonly proposed mathematical fairness notions. Potential advantages of these fairness notions are that they are explicit with clear boundary definitions. As such, they may be a way to very crisply define policy, and thus aid in making DEI policy more tangible, principled and evidence-based. 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.

Our research question therefore is: **How do organizational representatives understand and perceive different mathematical fairness notions in the context of early candidate selection in hiring?** We answer this question by qualitatively investigating our participants' conceptual understanding, perception of fairness, perception of diversity and judgment of applicability of using the different mathematical fairness notions during the early candidate selection in hiring. The organizational representatives comprise professionals in executive functions, talent acquisition, HR, organizational psychology, and diversity and inclusion operations.

Our study offers several contributions to the field. First, there still is much less empirical work on human perceptions and considerations of technical fairness notions and interventions, than there are proposals of new mathematical notions or interventions. The nascent field focusing on human perception of fairness is limited to lay people's, user's or ML designers understanding of various fairness scenarios [94, 89, 41, 58, 50, 27, 91, 62, 60]. With our work, we broaden the scope by engaging 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.

Second, within the hiring domain, our focus on early candidate selection is a novel take: much of existing literature focused on the automation of assessment instead. In our approach, we also purposefully will leave it 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 The hiring pipeline

The hiring pipeline of organizations is 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. 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, who select a smaller subset of applicants with each step. Finally, the most suitable applicants from the subset remaining in the last step are offered the jobs. Especially for the candidate selection stage, several data-driven algorithmic decision-making interventions have been proposed, offering automated candidate screening and assessment [19, 87, 13, 24]. Reasons for adopting such interventions involve overcoming unconscious human judgment [54], increasing the processing efficiency [12], and economic benefit [1].

2.2 Algorithmic systems for candidate selection

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 [63], [16]. A study [63] on interviewing HR Managers found that while some organizations onboard algorithmic tools after legal consultation, very few organizations undertake consultations on the type of data, input and target variables, validation processes or debiasing solutions employed.

This finding has implications on the organization's policies regarding hiring: without the ability to understand societal consequences of adopting the tool, organizations are unable to justify its use to the effect of fair and inclusive hiring.

2.3 Algorithmic fairness notions

Many fairness notions have been proposed based on data and sometimes, domain knowledge, which can be classified broadly as notions towards group fairness [23, 42, 8, 55, 21] or individual fairness [37, 30, 57, 52].

Group fairness, also known as statistical fairness, seeks to treat different social groups equally [57]. With minimum assumptions of the underlying population [51], it provides statistically-fair average guarantee for groups made of different sensitive attributes (e.g. race, education-level). For instance, Statistical Parity [30], requires equal prediction rates across groups defined by their sensitive feature. Other examples of group fairness notions include Equal Opportunity [42], Predictive Parity [21], Overall Accuracy Equality [8] and Calibration [21]. In contrast, individual fairness asks for similar predictions for similar individuals [57]. A fine-grained analysis of fairness can be achieved by placing constraints on pairs of individuals [30]. A motivation for desiring this granularity of fairness is that people who are less qualified should not be preferred over more qualified ones [48]. Examples of individual fairness notions include Causal Discrimination [37], Fairness through Awareness [30], Counterfactual Fairness [57] and No Unresolved Discrimination [52].

The mathematical implementations of different fairness notions can be targeted at different elements of the data-driven decision-making system pipeline, such as correcting the input data [49, 17], searching for a feasible solution space [42, 106, 108], representing the features as a graphical problem [67, 76, 52, 83], or using fair representation learning [53, 69, 25, 66]. The attempt to combine multiple fairness notions has been met with limitations, leading to treating individuals unfairly despite satisfying group fairness [9] or inability to mathematically satisfy multiple fairness notions at the same time [55, 74, 2]. Furthermore, the mathematical addressing of fairness has created concerns about the focus on trade-offs between accuracy and fairness [7, 106, 23, 32], while failing to address fairness from a societal point, creating a mismatch between what is required to be measured and it operationalization [47, 65].

2.4 Human-centered research on fairness notions

As the field of fairness in ML is growing, a large body of work has focused on human-centered research. To bridge the gap between the mathematical and social context of fairness in ML, several studies have been conducted to understand social perceptions of algorithmic decisions by lay people [94, 89, 41, 58, 10, 29, 90] and people affected by algorithmic decisions [105, 15, 103, 59].

In studies with lay people on several scenarios, people turned out to have difficulty judging fairness when presented with two notions and tend to prefer the simplest notion [94]. In high-stakes scenarios, people ended up preferring Statistical Parity as the most fair [94], while in case of lending, Calibration was the most preferred notion [89]. Some findings also suggest people's judgment depends on the severity and impact of decisions in fields such as recidivism [29], showing that people's concerns of fairness in recidivism went beyond topics of discrimination [41].

In the context of hiring, it was found that people perceive human decision to be more fair compared to an algorithmic decision, despite the decisions being the same [58]. This finding aligns with people's change in perception of justice with the change in human involvement in decision-making [10].

In situations where people are personally affected by algorithmic decision-making, it was found that they have low trust in the systems and negative emotions are evoked regarding racial and economic injustice [105, 15]. Moreover, people have favorability bias, where people prefer positive

judgments and negatively perceive them if a negative judgment was received [103, 59]. These findings can be attributed to a diverse set of human characteristics and demographics that impacts how people perceive fairness [103, 59].

2.5 Designer and user needs in designing fair algorithmic systems

Studies into lay peoples' perceptions on fairness show that social and human behaviors are difficult to mathematically express [59]. Recent literature has also looked into the needs of (potential) users or practitioners [91, 62, 60], and designers [50, 27, 87] of such systems.

On the side of the users, different fairness motivations exist based on their interaction with the system and the context in which they use the system [60]. The users prefer to choose systems aligned with their goals and values [91]. Many works in the human- computer interaction domain show how to embed social values into technical systems [56, 92, 93], but fairness as a value can cause confusing discussions in similar and different domains [75], which makes conceptualization of fairness difficult.

The designers need more support in undertaking interdisciplinary conversations and support in understanding the domain [27]. Moreover, they also request context specific guidance and the ability to identify the population composition of the users the systems are designed for [27, 44, 70].

This shows the disconnect between user and designer needs to produce practical and fair socio-technical systems [100], warranting further research. A promising direction towards participatory and collaborative approaches is now emerging in recent literature [68, 101, 61].

3 DESIGN OF TOY EXAMPLES

In our work, we will use fictional toy examples of early candidate selections, based on various mathematical fairness notions. In this section, we discuss how we chose the scoping of scenarios, which mathematical fairness notions we chose to adopt, and how these were visualized to our participants.

3.1 Scenario scoping

Before embarking on our main study, semi-structured interviews were first conducted with five organizational professionals, to better familiarize us with the problem domain, and understand how to scope our study such that it would be sufficiently recognizable and realistic. Following these interviews, several choices were made. First of all, we chose to frame scenarios as taking part in the earliest phase of candidate selection in the hiring pipeline. While later phases in the selection procedure may more explicitly focus on vacancy-specific differences between candidate suitability, and focus on advancing with very few people towards job offers, early stages have more focus on removing clearly unsuitable candidates, while leaving room for still advancing with potentially promising candidates, even though they may not be the most obvious match to a position. As such, there often is explicit room for advancing with a diverse pool of possible candidates and focus on aspects of fairness and broader DEI considerations.

In order to discuss scenarios in the context of fairness, a choice also had to be made on what sensitive attributes would need to be discussed. Our interviewees mentioned gender, nationality and ethnicity as important sensitive features. However, in current practice, gender is the only sensitive attribute that today is explicitly available and monitored by organizations to improve diversity, while other features in practice are not accessed or retained by organizations, also due to moral and legal dilemmas[1]. Therefore, for our study, we will focus on a scenario of gender

[1]Here, it is contextually relevant to mention that our study was conducted in the country of The Netherlands. In this country, since January 1, 2022, a law has entered into force defining an appointment quota for the Supervisory Boards of

imbalance. This is in line with the findings in [44], where designers do not have access to all the sensitive features but must work with coarse-level information.

While historical data and much of current data retained by organizations classifies gender in more categories than male and female, common mathematical fairness notions depart from a binary take to group membership. As a consequence, we presently opt for a binary representation of gender, and translate the gender imbalance scenario into toy examples where a particular job position historically had been male-dominated.

Finally, many possible mathematical definitions of fairness exist, that require different types of information. For example, some fairness definitions consider False Negatives (i.e., in the case of hiring, rejected candidates that in retrospect should have been selected). However, as per GDPR guidelines, applicant information cannot be retained by an organization without an applicant's consent, and our interviewees indicated that in practice was impossible for their organizations to monitor how rejected candidates were faring, after being rejected by the organization. Therefore, in our choices of fairness notions, we only could choose notions that would not consider False Negative rates.

3.2 Chosen mathematical fairness notions

For our study, we adopt 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. The 5 fairness notions are shown in Table 1.

Table 1. The chosen mathematical fairness notions

Fairness Type	Name	Formula	Citation
Group	Statistical Parity (SP)	$P(\hat{Y}\|A=0) = P(\hat{Y}\|A=1)$	[30]
Group	Equal Opportunity (EO)	$P(\hat{Y}=1\|Y=1,A=0) = P(\hat{Y}=1\|Y=1,A=1)$	[42]
Group	Calibration (CB)	$P(Y=1\|S=s,A=0) = P(Y=1\|S=s,A=1)\ \forall s \in [0,1]$	[21]
Individual	Fairness Through Awareness (FA)	$D(P(X_i),P(X_j)) \leq d(X_i,X_j)$	[30]
Individual	Counterfactual Fairness (CF)	$P(\hat{Y}_i\|A_i=0) = P(\hat{Y}_i\|A_i=1)$	[57]

listed companies, that should ensure that men and women each hold at least one third of the seats on the Supervisory Board. Furthermore, public and private limited liability companies are required to set appropriate and ambitious target ratios to improve the gender diversity on their boards and among their senior management personnel, and report on the current situation yearly [40]. At the same time, where in the Anglo-Saxon countries, the collection of ethnic/racial data has been common for the purpose of monitoring from an equality of opportunity perspective, such monitoring is formally absent in The Netherlands, as well as many more European countries [99].

3.2.1 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 [30]. 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 [4], where random candidates instead of qualified candidates are selected in order to satisfy statistical parity.

3.2.2 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 [42]. 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) [72].

3.2.3 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 : $P(Y = 1|S = s, A = 0) = P(Y = 1|S = s, A = 1)\ \forall s \in [0, 1]$ [21]. 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 $s \in [0, 1]$.

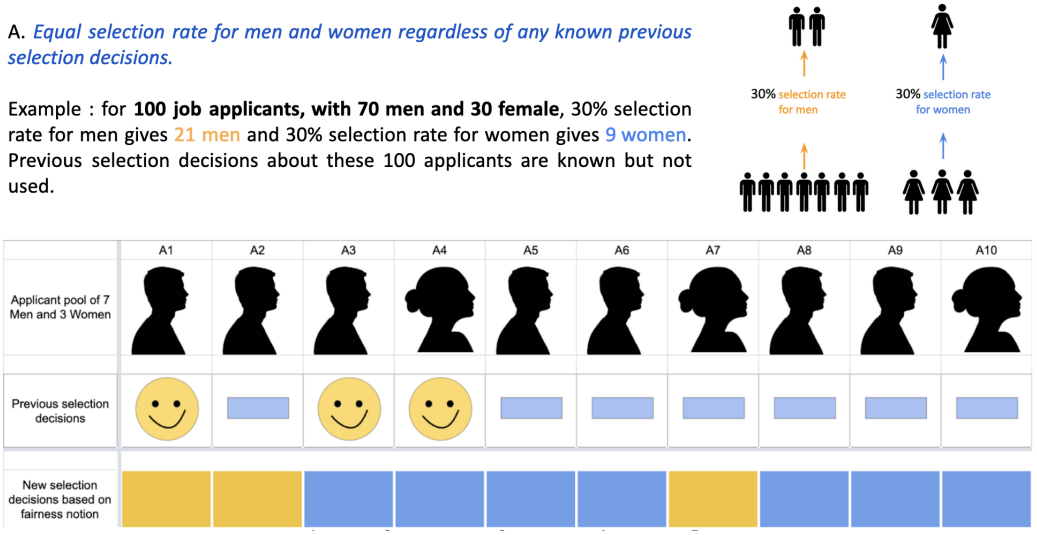
3.2.4 Fairness through awareness. Fairness through awareness (FA) requires the same prediction for any pair of individuals whose similarity falls under a given threshold [30]. For any two individuals i and j, where $P(X_i)$ is the probability distribution over all possible outcomes of prediction for 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.

3.2.5 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 [57]. 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.

3.3 Visualization of the fairness notions

(a) Initial design of the toy example for SP



(b) Final design of the toy example for SP

Fig. 1. Design changes to the toy example

While we strive to discuss perceptions of mathematical fairness notions with organizational representatives, it is unlikely that these representatives can easily comprehend the formal technical mathematical definitions of the fairness notions, and conceptualize what this would look like in an early candidate selection scenario. Therefore, we chose to visualize the chosen notions, inspired by the design used in [94]. However, where [94] displayed the outcomes of different algorithmic

strategies in parallel (and then found that participants preferred the most simplistic strategy), we will discuss one fairness notion at a time to avoid possible bias towards a simpler alternative.

In a direct translation of the design in [94], toy examples would look like in Figure 1a. However, in iterative consultation with several lab members (both with and without technical or mathematical backgrounds), we made several changes, leading to the design as shown in Figure 1b, with the full set of used designs being presented in Appendix A. As our visualizations were intended as discussion-starters rather than fully optimized designs, we did not go through a full iterative design and development process with a broader user population yet, and thus cannot yet make formal claims on accessibility and user-friendliness of these visualizations. Still, several relevant reasonings with regard to our design choices are discussed below.

First, 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 make the selections. As discussed in Section 2.4, people may also have different perceptions of trust in outcomes following from algorithmic or human procedures. To avoid such bias, we removed any reference to algorithmic procedures. Instead, the fairness notion of interest are contrasted with 'what a previous selection committee' would have judged. With this change, we avoid potential response bias, as the type of decision-maker for the same task evokes different emotions in people [58].

Next, we found the concept of 'true outcome' problematic in the context of hiring procedures: it could imply an absolute ground truth that a prediction should match closely to (which 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 true outcome 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.

Similarly, the concepts of 'selected' and 'rejected' appeared too harsh, implying 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. Therefore, we chose to visually give a positive association to those who would previously get selected, but a neutral indication for those who would not. Additionally, 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. yellow palette.

Our conversations with independent colleagues (in particular, the person most remote to technical work) also informed our chosen order of discussion for the different fairness notions. In line with the finding in [94], SP was easily understood. EO was interpreted as an 'improvement' on SP, so discussing SP before EO when discussing group fairness would be logical, while CB appeared hardest to understand. As for the individual fairness notions, FA triggered less clarification discussions than CF, and thus would be easier to discuss first.

4 METHODOLOGICAL SETUP

4.1 Recruitment

We used a combination of purposive sampling and snowball sampling [39] 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 personal connections in industry. Next, we made announcements on LinkedIn inviting professionals in the same domains 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

participate in an interview by sending us an e-mail. We also found a list of professionals working in this field using LinkedIn, whom we contacted by sending direct messages.

We recruited experts working in The Netherlands between May and July of 2022, who had professional experience regarding topics of diversity, equity and inclusion in the hiring domain. Being aware that the HR field tends to largely feature white women, we also made explicit efforts towards recruiting a diverse participant group through active LinkedIn searches. From the lists of professionals found on this platform, we prioritized reaching out to those in minority groups.

Through these strategies, we were able to find and contact a little over 80 potential participants, out of whom 48 responded, and 21 agreed to schedule interviews within the time-frame of the study. We conducted the study at Delft University of Technology, The Netherlands.

4.2 Participants

Out of the 21 participants, we selected 17 for our more thorough analysis[2]. The professional roles of all participants were verified with their Linkedin profile and their demographic information is listed in Table 2. Reflecting our conscious effort to obtain a diverse participant group, our sample consists of 9 women and 8 men, of whom 5 self-reported as belonging to an ethnic minority, 4 self-reported as belonging to other minority related to age, health, sexual orientation, immigration and neurodiversity, and only 1 identified as belonging to both an ethnic and another minority group. All participants work in The Netherlands, at a diverse set of differently-sized organizations, as shown in Table 3.

4.3 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 be collected, purpose of audio recording, participant's right to withdraw from the research and contact details of the researchers. After

[2]We excluded 4 participants from the analysis, as in the interviews, they turned out to have insufficient hands-on experience in HR and DEI initiatives.

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

obtaining the participant's explicit and voluntary consent to the opening statement, we proceeded with the semi-structured interviews.

The interview consisted of 6 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 phase in candidate selection, examples of job positions such as engineers, nurses or accountants and possibility of selecting multiple candidates for further stages in the hiring pipeline.
(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.
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), in line with increasing difficulty as perceived when discussing our designs (Section 3).
(4) The discussion of each fairness notion concluded with 5-point Likert scale responses 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)
(5) 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.
(6) Finally, participants were asked to describe their experience of the interview and their thoughts about using the fairness notions.

4.4 Analysis

Audio recordings of the interviews (totaling 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 for coding. We employed a combination of inductive and deductive coding [31, 85] to conduct a thematic analysis [14] of the interview transcripts.

Each of the 17 interview transcripts were divided into 5 sections corresponding to the 5 fairness notions. Initially, both authors of this paper followed an inductive approach and independently open-coded four randomly chosen interview transcripts at sentence and paragraph level to create codes while analyzing the transcripts. The goal of this preliminary analysis was to create a set of codes by two coders, which would later be discussed and revised to establish a reference codebook. The sample size of 4 transcripts (23.9%) in the preliminary analysis is a sufficient proportion to be coded, which is recommended to be between (10-25%) [80]. Each fairness notion was coded independently, giving codes corresponding to each fairness notion. Coding 4 interviews this way produced 2 sets of codes by two authors for each of the 5 fairness notions.

Following this preliminary analysis, iterative discussions were held between the two authors to compare, identify and consolidate the codes as different coders interpret and organize their codes differently [3].The discussions were guided by the research question to establish the reference codebook for the next part of the analysis. Discussion is an acknowledged form of inter-coder agreement [22] which improves the reliability of the codes [45]. Quantitative measures of inter-coder reliability prior to the discussions were not necessary, as the two coders jointly discussed and revised the codebook, leading to alignment between the coders.

For each of the 5 fairness notions, the codebook contains the participant ID, codes given by the authors to the sentences in the transcript and the frequency of occurrence of each code across transcripts. For instance, participant P8's quote when talking about Statistical Parity, *"No, the selection rate needs to go up. You need to identify your minority group and make sure it is at least 50% of your talent pool. Otherwise you'll never make this change."*, was assigned the high-level code of 'diversity', followed by the sub-code of 'talks about minority' and 'can't help'. The same sentences were further coded with 'suggestion for improvement'.

As one coder is sufficient to code all the transcripts after establishing a reference codebook [18], the first author open-coded all the 17 transcripts by applying the codes from the codebook following a deductive approach. A code was assigned to multiple sentences or paragraphs containing similar information. As the codebook was established based on a smaller subset of transcripts, additional relevant codes were applied to the transcripts using an inductive approach to cover concepts not identified in the reference codebook.

Thereafter, the codes were categorized into high-level codes and sub-codes, and grouped per fairness notion. The categorization of codes this way produced themes on conceptual understanding, perception of fairness, perception of diversity, applicability and concerns of participants regarding the 5 different mathematical fairness notions during the early candidate selection in hiring. The themes are described in detail in section 5. In conducting the interviews, it was empirically observed that the responses given by the participants became predictable after 12 interviews. Moreover, very few new codes emerged from the last two interviews. This suggests that saturation was reached, and the sample size of 17 participants in total was justifiable.

For each fairness notion, we collected Likert-scale ratings from our participants, to allow for more structured visualizations of how participants rated the 3 components of understanding, fairness

and diversity. Furthermore, this allows us for grouping rating distributions for different participant segments (i.e, considering the respondent's gender, or self-identified membership of an ethnic or other minority). Finally, through alluvial plots, we visualize whether and how ratings for each of the notions changed after the participants discussed all notions.

Given our small sample size, these visualizations should be seen as exploratory illustrations, more than true quantitative analyses. We will therefore also refrain from drawing formal statistical conclusions from the rating data.

4.5 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 RESULTS

In reporting our results, we visualize the ratings given by participants on Understanding, Fairness and Diversity, together with breakdowns for sensitive self-reported features (Gender, belonging to an Ethnic Minority, belonging to an Other Minority). Following our thematic analysis on the interview transcripts, for each fairness notion, we 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.

5.1 Statistical Parity (SP)

Coding for SP produced 150 quotations. Rating distributions of the participants are shown in Figure 2.

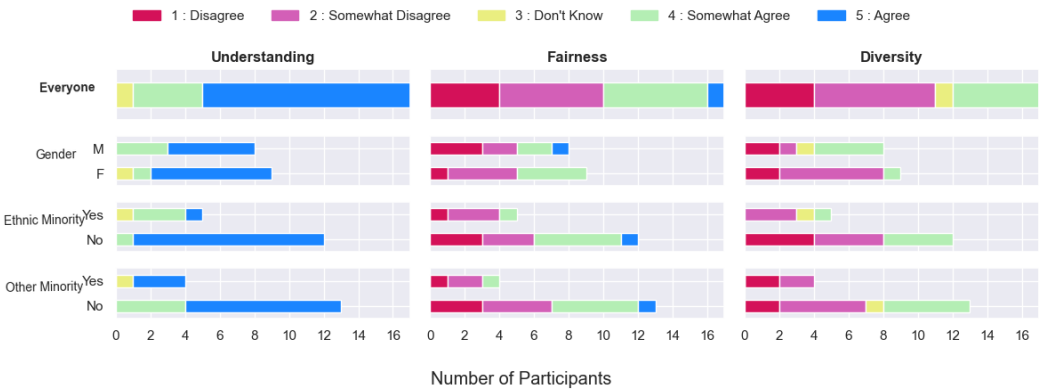


Fig. 2. **SP** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of SP, along with the distribution by *gender, ethnic minority* and *other minority* groups.

5.1.1 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. Several 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, several participants also disagreed with the notion refusing to 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.

Some 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.

5.1.2 Perception on fairness of using SP. Majority of the 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, many 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, few 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 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. Some 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.

However, a larger number of participants, reasoned about the notion's inability to improve diversity for two main reasons (E2, E3, E5, E8, E12, E13, E15, E16). First, some 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, another set of participants 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, some 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.

5.1.4 Applicability of SP. The majority of participants were concerned about the structure of the selection rate. They 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. Some of them were concerned about gender, rather than 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, some 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.

5.2 Equal Opportunity (EO)

Coding for EO produced 88 quotations. Rating distributions of the participants are shown in Figure 3.

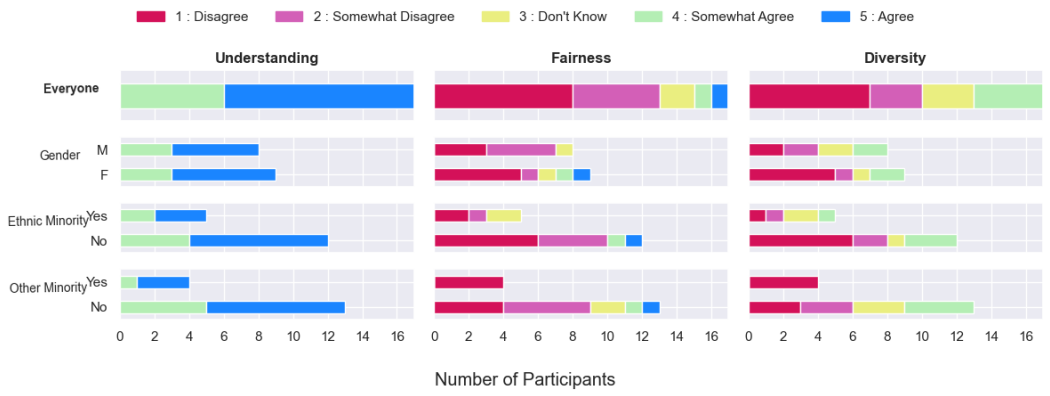


Fig. 3. **EO** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of EO, along with the distribution by *gender, ethnic minority* and *other minority* groups.

5.2.1 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, many 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. Several 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 interfere with your selection method knowing what somebody else concluded."* - E1. *"If you ask me what happened and why did they do this? That part I don't understand."* - E17.

5.2.2 Perception on fairness of using EO. A majority of the 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. Several participants expressed the notion being unfair due to the influence of actual outcome in the selections and the need for more information on reasons for actual outcome (E1, E2, E4, E7, E10, E15, E17). *"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 few 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. *"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 Perception on improving diversity when using EO. We see that many participants were certain that using EO will not improve gender balance in organizations. Many of them attributed this to the 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.

5.2.4 Applicability of EO. The biggest source of concern for the majority of the participants was trusting the actual outcome. Many 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. *"As an organization we never ask the opinion of the previous committee."* - E10.

While majority of the participants expressed concerns about using the actual outcome, some participants mentioned that they would only use the notions if the reasons behind the actual outcome is known (E6, E7, E9, E16). *"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. *"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. Many 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, it was also indicated 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)

Coding for CB produced 72 quotations. Rating distributions of the participants are shown in Figure 4.

5.3.1 Impression of CB. Most participants said that they found CB 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't really find it"* - E17. *"I think this system is really confusing. 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, few 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. Some 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, some people also disliked comparing applicants and assigning them scores (E4, E10).

5.3.2 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. Few participants indicated that they needed more information and context (E13, E17). *"I can't really*

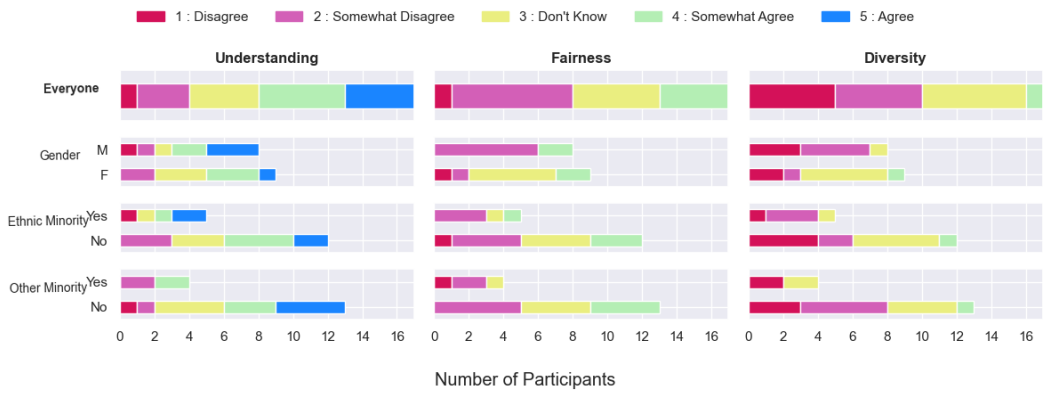


Fig. 4. **CB** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of CB, along with the distribution by *gender, ethnic minority* and *other minority* groups.

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. Few 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.

5.3.3 Perception on improving diversity when using CB. The majority of the participants rated CB 3 or lower for diversity, which may relate to this notion having been harder to grasp. Several 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 Applicability of CB. Participants expressed two major concerns regarding the applicability of CB. First, 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 (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)

Coding for FA produced 127 quotations. 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.

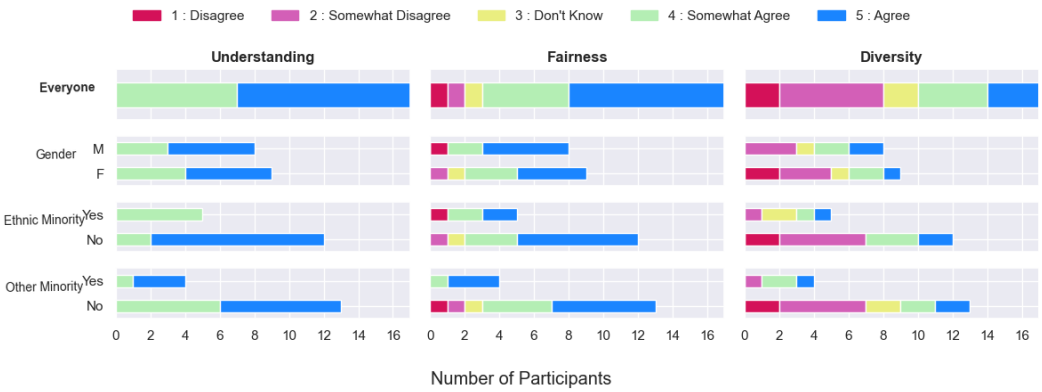


Fig. 5. **FA** - Ratings given by participants on *Understanding, Fairness* and *Diversity* of FA, along with the distribution by *gender, ethnic minority* and *other minority* groups.

5.4.1 Impression of FA. Participants had many things to say about FA. While several participants agreed with the notion saying they were happy that both genders were treated equally (E2, E4, E5, E6, E9, E17), few 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 many 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's where you want to go as an end goal but to drive the change for unconscious bias, you need to create more opportunities for women"* - E8.

5.4.2 Perception on fairness of using FA. A 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 some 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, some participants found it quite difficult to rate fairness of FA saying that fairness depends on many factors (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 Perception on improving diversity when using FA. In contrast to ratings on fairness, ratings on diversity for FA are divided. While some participants said that FA could help organizations improve gender balance (E4, E5, E15), many others 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, some 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.

5.4.4 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, several 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)

Coding for CF produced 72 quotations. Rating distributions of the participants are shown in Figure **??**.

5.5.1 Impression of CF. CF was very well understood by the majority of the participants. 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 they are ethically speaking the same, right?"* - E7.

5.5.2 Perception on fairness of using CF. Perception of fairness and diversity of using CF elicited few responses from the participants. Some 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. While few other 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 Perception on improving diversity of using CF. A majority of the participants was unsure whether CF would help improve the gender balance in organizations and indicated 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 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 Applicability of CF. The biggest concern about CF expressed by participants was related to gender. Some 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, 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.

It was also indicated that CF has low applicability, when diversity is a goal of the company (E7, E11). On the applicability of CF, some 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 through many cycles to reach true fairness levels. So it's not always applicable."* - E5. It was also mentioned 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 Changes in the ratings

Having discussed and rated all the notions, participants were asked to rate the same notions on Fairness 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 6 visualizes to what extent initial and final ratings by participants changed for each fairness notion. Again, as the study is an interview study, the ratings were obtained for exploratory illustrations.

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.

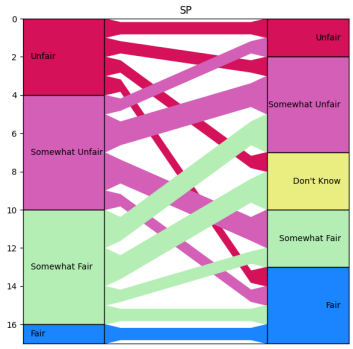
6 DISCUSSION

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.

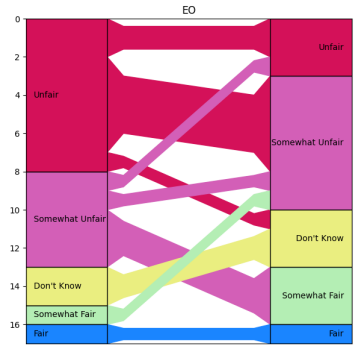
6.1 No fairness notion emerged as the most suitable for early candidate selection in hiring

Where overall, participants had high understanding of the fairness notion explanations, Calibration (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 [94], 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 Statistical Parity conflicting with considerations of merit, which are important in hiring, where [94] considered other application domains (recidivism prediction and skin cancer diagnosis), while the raters were no domain experts. At the same time, in our case, Statistical Parity did not stand out as obviously simpler to understand in comparison to other notions.

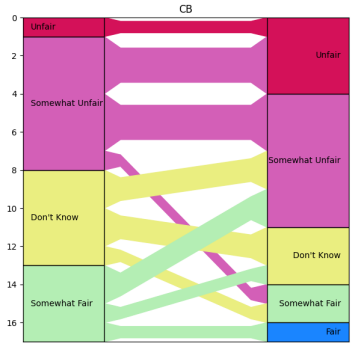
Despite a high degree of understanding on most fairness notions, most participants found it quite difficult to provide a rating for fairness as they did not have a clear set definition for fairness themselves. Most participants provided their rating after talking about multiple dimensions of fairness, asking whether it was fair to organizations or applicants. This is in line with earlier findings that definitions of fairness can differ within a domain and cause confusion [75]. This points to contextualizing and defining the scope of fairness in future studies, as people find it quite subjective. 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 Fairness Through Awareness and Counterfactual Fairness and rated it as fair. However, their language suggested some restrain or skepticism. 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. This is aligned with findings from [41], where participants had no consensus on fairness and their discussion went beyond discrimination. In our work, the main theme surrounding the reasoning



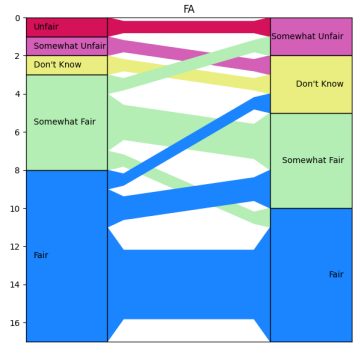
(a) Statistical Parity (SP)



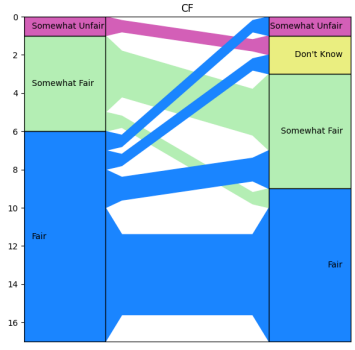
(b) Equal Opportunity (EO)



(c) Calibration (CB)



(d) Fairness Through Awareness (FA)



(e) Counterfactual Fairness (CF)

Fig. 6. Alluvial plot showing changes in the ratings for all fairness notions. Initial ratings on the left move to final ratings on the right.

behind low ratings for fairness and diversity was missing context: 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. Here, it is likely that a 'one size fits all' approach will not exist. Zooming out of the need to choose an existing mathematical notion, it can be argued that our participants' hesitance in advocating for most notions comes from their discomfort in choosing distributive nature of fairness, which most notions are based on. Formulations based on other forms of justice, for instance relational equality [33] are only recently being researched and could incorporate aspects that the current notions lack. It is likely that algorithmic decisions could only support human work, rather than replace it. Thus, more effort will need to be spent on technical addressing of fairness where designers request context specific guidance [27], users define their fairness motivations depending on the context of the interaction with systems [60] and experts, as seen in our work provide context on domain needs.

6.2 Fairness and Diversity considerations may not go hand in hand in the hiring domain

The complex relationship between fairness and diversity emerged with higher ratings for fairness but lower ratings for diversity for most notions. 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. Similarly, as already discussed in the reactions to Fairness Through Awareness and Counterfactual Fairness, what is considered fair may not be applicable or defensible in practice. This points towards a need to look at holistic view of fairness and diversity because the fairness notions may not be sufficiently rich to capture societal needs regarding diversity. This aligns with literature suggesting looking beyond discrimination [41] because different scopes impact how people perceive fairness [58]. In our study, we uncover that diversity is an important consideration when addressing fairness in hiring. The major concern noted for Statistical Parity, Equal Opportunity, Fairness Through Awareness and Counterfactual Fairness 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 Statistical Parity and Equal Opportunity, it was in relation to every individual being defined by different characteristics for Fairness Through Awareness and Counterfactual Fairness. Moreover, participants expressed that a skewed applicant pool with fewer or no minorities makes it difficult to put notions such as Statistical Parity and Equal Opportunity into practice. The applicability of fairness notions, though used in practice [87], is under-researched in literature, pointing towards the need for collaboration between different stakeholders when designing fairness promoting directives in organizations. With our work, we show that organizational representatives in various functions such as HR, DEI and I/O psychology hold key knowledge in shaping hiring policies, and thus have valuable insights to share in discussions on the possible adoption of mathematical fairness notions towards fair hiring.

6.3 Concerns of organizational representatives on discussing topics of fairness

A recurring theme while discussing fairness notions with the participants was that our depictions were 'too simplified' a concept, while fairness in itself encompasses many complex dimensions that are difficult to mathematically represent. Many participants were hesitant to provide a rating, had difficulty explaining their ratings or had to be asked multiple times to choose the most suitable rating before moving to the next question, especially on their perception of fairness. Their main concern arose from their rating not able to fully justify societal and organizational needs. This highlights the directions for future inter-disciplinary conversation: tools to aid conversations

surrounding topics of fairness could allow for the expression of varied dimensions of fairness to gather more holistic insights.

One such source of discomfort among participants of this work was whether to include a sensitive feature such as gender so explicitly in the selection process. This emerged from the fear of selecting unqualified female candidates for the 'sake' of fairness. Many participants said that merit was important and diversity considerations should not precede merit. 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. While it is known that in machine learning methods, removing the sensitive attributes from the input features does not remove discrimination because of their correlation with other attributes, this could not be explicitly discussed with our participants as they were not made aware of the mathematical origin of the fairness notions. Previous studies [41, 89, 90] have shown that lay people question the use of sensitive features in algorithmic decision making and its usage lowers their trust in such systems. Further, a handful of participants also showed discomfort in the binary treatment of gender, asking to include more options. The discomfort in weighing an attribute such as gender in hiring despite organizational policies surrounding diversity indicates the delicate nature of such a topic and the organization's need to tread such conversations with caution.

Lastly, participants expressed concern regarding the use of actual outcome in making current decisions, with most of the participants questioning its use and validity in the current decisions. Specifically, for Equal Opportunity, participants said that they need space to disagree with the actual outcome because it could perpetuate bias from previous decisions. However, when reflecting on Calibration, the usage of the actual outcome brought concerns about the lack of transparency in decision making. There are two main implications of this finding. First, as actual outcome comes from the input data for algorithmic systems, this highlights concerns about data quality. It is possible that clarity and transparency in data sourcing methods could improve the organizational representatives' trust in using the actual outcome. With that respect, concerns exist that aspects of algorithmic systems are not often disclosed fully by vendors [87]. However, a study [103] with lay people has shown that final outcomes are more important that the process. This contrast can be attributed to the fact that in [103], the outcomes personally affected the lay people whereas, in our work, the participants are domain experts. With ramifications on diversity goals of organizations, transparency is needed on what data is collected and what is measured because it can affect the users' trust in the system [90]. Second, while mathematical fairness notions are designed with the view of distributive justice, our participants also expressed concerns about procedural justice. The difficulty in trusting the use of actual outcome, the need to understand the concept of the fairness notions and concern about use of sensitive features to make new selection decisions shows that organizations want to recognize and justify the use of selection procedures. The plethora of literature in [88] shows that AI and fairness are increasingly being discussed in organizations, through the lens of some forms of justice. However, these discussions have limited scope and do not fully incorporate concepts such as employee engagement, trust, or socio-economic impact, pointing to the need of broader inter-disciplinary perspectives.

In terms of future adoption of technically enabled decision making, the above considerations suggest that any algorithmic approach will need to be positioned in spaces with sufficient room for ongoing discussion and reflection by different human stakeholders. As such, algorithmic data-driven systems could be seen as digital support to a human process, where the decision-making aspects strongly need to remain on the human side. The formulation of the current fairness notions, as seen with discussion with our participants, lack context and human involvement.

6.4 Minoritized groups may bring specific, different stances on how to improve diversity

In [105], it is shown how algorithmic fairness evokes negative emotions from minority groups regarding racial and economic justice. While our sample has been too small to draw strong conclusions, as most of our current participants did not identify as belonging to a minority, 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. 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. A similar result, albeit in a different context of course recommendation and recidivism shows that men are more likely to choose maximizing accuracy over minimizing racial disparities and they also prefer algorithms over human judgment [86].

6.5 Fairness notions are no all-encompassing solution 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 would be needed for fair and diverse hiring. For instance, many participants expressed the need to improve the preceding and succeeding stages to early candidate selection. These stages involve attracting suitable candidates to apply and removing human biases that can appear in interview stages. With our work, it becomes clear that interventions for fair hiring needs to be considered at multiple stages of the hiring pipeline, thereby deeming the fairness notion as only a support in one important phase of this pipeline.

Our interviews demanded considerable time investment from our participants (60-75 minutes), sometimes requiring a second sitting to complete the session. While no resources were available to compensate participants for their time, the 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 stakeholders are eager to remain involved. Considering the need for human-centered, interdisciplinary and holistic perspectives in which the technical solution would be a supporting element in larger decision-making procedures, we believe this makes future interdisciplinary collaborations realistic, and would expect for these to particularly be a good fit to the CSCW and HCI research communities.

6.6 Limitations and Future Work

Several limitations can be identified in our current work. First of all, while our non-technical translation visualizations of mathematical fairness notions were developed with feedback from independent colleagues, they have not formally been evaluated for user-friendliness and accessibility. In order to turn them into a contribution that other researchers can confidently build upon, more thorough study will need to be done with regard to their design, with further design iterations, that should more structurally be tested with broader audiences..

Continuing on discussions on multifacetedness, our current work also only considered one sensitive attribute (gender) and treated it as a binary variable. 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 including multi-valued sensitive attributes, and aspects of intersectionality, which may lead to extra adverse effects on those belonging to multiple minoritized groups [16]. Next to this, it will be more realistic to not only depart from binary variable membership (such as a binary take on gender), but allow for multiple possible categories or values within a variable.

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 but many mathematical notions cannot be satisfied simultaneously [2, 21, 55]. 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.

Generally, the question of fairness needs to be considered in the context of its application. Even within the scope of early candidate selection, 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.

Choices of fairness notions reflect choices of world views, and challenges of untangling and balancing these. Different stakeholders and people coming from different background experiences will bring different perspectives. In our current work, we did not address this as explicitly yet, while for future work, it is important to address this more thoroughly. For example, 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. It e.g. 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 [9]. Thus, what our respondents currently prefer may not actually be the preferred course of action. As being rejected in a job selection process could be seen as being denied an opportunity for financial stability, this suggests a distinction between social acceptance and ethical acceptability. In the field of Ethics, for other dilemmas on the adoption of risky technology, arguments were made that both social acceptance and ethical acceptability need to be considered when policy is to be set [95]. This 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 [96].

Furthermore, as we discussed, it appears that participants representing minorities may have different reactions to fairness and DEI policies than those representing majorities. We hypothesize there may be different reasons for this. For example, those from minorities may themselves have faced discrimination. Furthermore, in situations of bias due to social inequality, it has been noted that those in power positions tend to not be representative of those who may be disadvantaged [73]. Continuing the discussion on power dynamics, fairness policy has been criticized for largely representing the interests of managers in an organization, which again may represent a status quo that is not necessarily mindful of the interest of minorities [35].

While we plead for more explicit inclusion of minority stances in future work, this does raise concern on how the right voices and stances can indeed sufficiently be heard without getting over-burdened. First of all, for future studies, we recommend to more explicitly check on concrete and lived experience of participants with aspects of discrimination, as well as the underlying intrinsic motivations of participants to participate in studies like ours. Secondly, to inclusively hear

the voices of minorities, it is important to situate problems in contexts they indeed relate to [46], which may imply that the current, more organizational, perspective of who gets to be hired may need to be reversed, and rather focus on what it means not to be hired or prioritized.

7 CONCLUSION

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.

When starting the research leading to this paper, initially, the authors had intended to work on (quantitative) fairness monitoring tooling for early selection stages in hiring. 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.

8 ACKNOWLEDGEMENTS

We thank Han-yin Huang, Linnet Taylor, Jaehun Kim for their insights and support to enrich this study. We extend our gratitude to various members of the Computer Science Department at TU Delft for extended discussions on our design choices.

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] David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. 1997. The place of inter-rater reliability in qualitative research: an empirical study. *Sociology*, 31, 3, 597–606.

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

[5] 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.

[6] Ruha Benjamin. 2019. *Race After Technology*. Wiley.

[7] 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.

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in criminal justice risk assessments: the state of the art. *arXiv:1703.09207 [stat]*, (May 27, 2017). Retrieved Mar. 7, 2022 from http://arxiv.org/abs/1703.09207 arXiv: 1703.09207.

[9] 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.

[10] 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.

[11] Sarah Bird, Miro Dudík, 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/.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21, 2, (Sept. 2010), 277–292. DOI: 10.1007/s10618-010-0190-x.

[18] John L Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. 2013. Coding in-depth semistructured interviews: problems of unitization and intercoder reliability and agreement. *Sociological methods & research*, 42, 3, 294–320.

[19] 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.

[20] [n. d.] Chief diversity officer appointments continue surge in 2022. https://businesschief.com/sustainability/chief-diversity-officer-appointments-continue-surge-in-2022. Accessed 24-12-2022. ().

[21] 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.

[22] Sherice N Clarke, S Sushil, Katherine Dennis, Ung-Sang Lee, Andrea Gomoll, and Zaynab Gates. 2023. Developing shared ways of seeing data: the perils and possibilities of achieving intercoder agreement. *International Journal of Qualitative Methods*, 22, 16094069231160973.

[23] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

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

[25] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*. PMLR, 1436–1445.

[26] Catherine d'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. The MIT Press.

[27] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency. ACM, Seoul Republic of Korea, (June 21, 2022), 473–484. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533113.

[28] [n. d.] Diversity wins: how inclusion matters. https://www.mckinsey.com/featured-insights/diversity-and-inclusion/diversity-wins-how-inclusion-matters. Accessed 24-12-2022. ().

[29] 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.

[30] 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.

[31] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5, 1, 80–92.

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

[33] Benjamin Fish and Luke Stark. 2022. It's not fairness, and it's not fair: the failure of distributional equality and the promise of relational equality in complete-information hiring games. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '22: Equity and Access in Algorithms, Mechanisms, and Optimization. ACM, Arlington VA USA, (Oct. 6, 2022), 1–15. ISBN: 978-1-4503-9477-2. DOI: 10.1145/3551624.3555296.

[34] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. 2016. 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. *Fed. Probation*, 80, 38.

[35] Marion Fortin and Martin R. Fellenz. 2008. Hypocrisies of fairness: towards a more reflexive ethical base in organizational justice research and practice. *Journal of Business Ethics*, 78, 415–433.

[36] 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.

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

[38] [n. d.] Getting serious about diversity: enough already with the business case. https://hbr.org/2020/11/getting-serious-about-diversity-enough-already-with-the-business-case. Accessed 24-12-2022. ().

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

[40] Government of The Netherlands. 2021. New legislation will improve gender diversity on corporate boards. (Sept. 2021). https://www.government.nl/latest/news/2021/09/29/new-legislation-will-improve-gender-diversity-on-corporate-boards.

[41] 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.

[42] 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.

[43] Annemarie M.F. Hiemstra, Eva Derous, Alec W. Serlie, and Marise P. Born. 2012. Ethnicity effects in graduates' résumé content. *Applied psychology*.

[44] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: what do industry practitioners need? 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), 1–16. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300830.

[45] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. 2004. Reliability in coding open-ended data: lessons learned from hiv behavioral research. *Field methods*, 16, 3, 307–331.

[46] Han-Yin Huang and Cynthia C. S. Liem. 2022. Social Inclusion in Curated Contexts: Insights from Museum Practices. (May 10, 2022). arXiv: 2205.05192[cs]. DOI: 10.1145/3531146.3533095.

[47] 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.

[48] 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.

[49] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. 2009 2nd International Conference on Computer, Control and Communication (IC$). IEEE, Karachi, Pakistan, (Feb. 2009), 1–6. ISBN: 978-1-4244-3313-1. DOI: 10.1109/IC4.2009.4909197.

[50] Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. 2021. 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. ACM, Virtual Event Canada, (Mar. 3, 2021), 690–700. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445931.

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

[52] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

[53] Diederik P Kingma and Max Welling. 2014. Stochastic gradient vb and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*. Vol. 19, 121.

[54] 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.

[55] 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.

[56] Cory Knobel and Geoffrey C Bowker. 2011. Values in design. *Communications of the ACM*, 54, 7, 26–28.

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

[58] 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.

[59] Min Kyung Lee and Su Baykal. 2017. 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. ACM, Portland Oregon USA, (Feb. 25, 2017), 1035–1048. ISBN: 978-1-4503-4335-0. DOI: 10.1145/2998181.2998230.

[60] 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.

[61] Min Kyung Lee et al. 2019. Webuildai: participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3, CSCW, 1–35.

[62] 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*. AIES '21: AAAI/ACM Conference on AI, Ethics, and Society. ACM, Virtual Event USA, (July 21, 2021), 166–176. ISBN: 978-1-4503-8473-5. DOI: 10.1145/3461702.3462531.

[63] 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.

[64] 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.

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

[66] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*. PMLR, 4114–4124.

[67] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv:1805.05859 [cs]*, (May 15, 2018). Retrieved May 5, 2022 from http://arxiv.org/abs/1805.05859 arXiv: 1805.05859.

[68] Daria Loi, Christine T Wolf, Jeanette L Blomberg, Raphael Arar, and Margot Brereton. 2019. Co-designing ai futures: integrating ai ethics, social computing, and design. In *Companion publication of the 2019 on designing interactive systems conference 2019 companion*, 381–384.

[69] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2017. The variational fair autoencoder. *arXiv:1511.00830 [cs, stat]*, (Aug. 9, 2017). Retrieved May 4, 2022 from http://arxiv.org/abs/1511.00830 arXiv: 1511.00830.

[70] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the fairness of ai systems: ai practitioners' processes, challenges, and needs for support. *Proc. ACM Hum.-Comput. Interact.*, 6, CSCW1, Article 52, (Apr. 2022), 26 pages. DOI: 10.1145/3512899.

[71] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20: CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, (Apr. 21, 2020), 1–14. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376445.

[72] 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.

[73] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying up machine learning data: why talk about bias when we mean power? In *Proceedings of the ACM on Human-Computer Interaction*. Vol. 6.

[74] 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.

[75] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This thing called fairness: disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, 3, CSCW, 1–36.

[76] Razieh Nabi and Ilya Shpitser. [n. d.] Fair inference on outcomes, 10.

[77] 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.

[78] Safiya Umoja Noble. 2018. *Algorithms of Oppression*. NYU Press.

[79] Cathy O'Neil. 2016. *Weapons of Math Destruction*. Crown Books.

[80] Cliodhna O'Connor and Helene Joffe. 2020. Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods*, 19, 1609406919899220.

[81] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19: Conference on Fairness, Accountability, and Transparency. ACM, Atlanta GA USA, (Jan. 29, 2019), 39–48. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287567.

[82] Samir Passi and Steven J Jackson. 2018. Trust in data science: collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction*, 2, CSCW, 1–28.

[83] Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2.

[84] Andrea Pemberton and Jennifer Kisamore. 2023. Assessing burnout in diversity and inclusion professionals. *Equality, Diversity and Inclusion*, 42, 1, 38–52.

[85] Chad Perry and Oystein Jensen. 2001. Approaches to combining induction and deduction in one research study. In *Conference of the Australian and New Zealand Marketing Academy, Auckland, New Zealand*.

[86] Emma Pierson. 2018. Demographics and discussion influence views on algorithmic fairness. (Mar. 4, 2018). Retrieved Dec. 23, 2022 from http://arxiv.org/abs/1712.09124 arXiv: 1712.09124[cs].

[87] 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.

[88] Lionel P Robert, Casey Pierce, Liz Marquis, Sangmi Kim, and Rasha Alahmad. 2020. Designing fair ai for managing employees in organizations: a review, critique, and design agenda. *Human–Computer Interaction*, 35, 5-6, 545–575.

[89] 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.

[90] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "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. ACM, Seoul Republic of Korea, (June 21, 2022), 1616–1628. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533218.

[91] Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. 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. ACM, Seoul Republic of Korea, (June 21, 2022), 440–451. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533110.

[92] Katie Shilton et al. 2018. Values and ethics in human-computer interaction. *Foundations and Trends® in Human–Computer Interaction*, 12, 2, 107–171.

[93] Katie Shilton, Jes A Koepfler, and Kenneth R Fleischmann. 2014. How to see values in social computing: methods for studying values dimensions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 426–435.

[94] 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.

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

[96] 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.

[97] Christine Teelken and Karin Kee. 2023. Male 'play-garden' versus female 'tightrope walking': an exploration of gendered embodiment in dutch higher education. *Studies in Higher Education*, 0, 0, 1–14. DOI: 10.1080/03075079.2023.2208158.

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

[99] Sharon van Geldere, Rozemarijn Stadens, and Linnet Taylor. 2022. *Anti-discrimination data collection in academia: an exploration of survey methodology practices outside of The Netherlands*. The Young Academy, Amsterdam.

[100] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. 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: 10.1145/3173574.3174014.

[101] John Vines, Rachel Clarke, Peter Wright, John McCarthy, and Patrick Olivier. 2013. Configuring participation: on how we involve people in design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 429–438.

[102] Sara Wachter-Boettcher. 2017. Ai recruiting tools do not eliminate bias. *Time Magazine*.

[103] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making algorithm outcomes, development procedures, and individual differences, 14.

[104] Gregor Wolbring and Aspen Lillywhite. 2023. Burnout through the lenses of equity/equality, diversity and inclusion and disabled people: a scoping review. *Societies*, 13, 5. DOI: 10.3390/soc13050131.

[105] 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.

[106] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. *arXiv:1702.06081 [cs]*, (Nov. 1, 2017). Retrieved Jan. 19, 2022 from http://arxiv.org/abs/1702.06081 arXiv: 1702.06081.

[107] Janice D. Yoder. 1991. Rethinking tokenism: looking beyond numbers. *Gender & Society*, 5, 2.

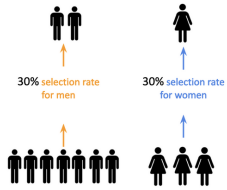
[108] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. *Proceedings of the 26th International Conference on World Wide Web*, (Apr. 3, 2017), 1171–1180. arXiv: 1610.08452. DOI: 10.1145/3038912.3052660.

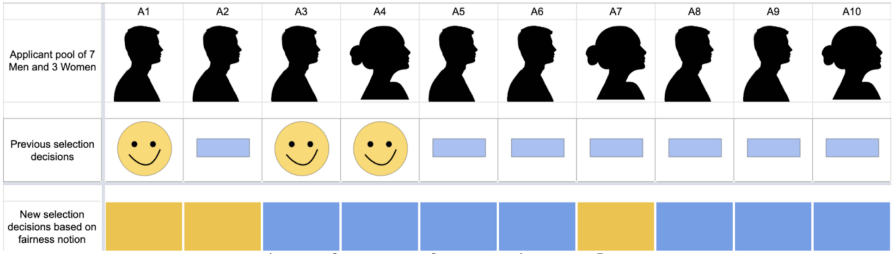
A FINAL DESIGNS

Received January 2023; revised July 2023; accepted November 2023

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.





(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.

(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.

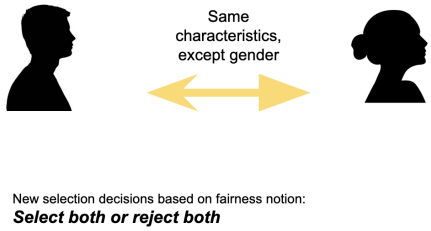
Give points by making use of previous selections.

Consider chance of previous selection or rejection

(a) Final design of the toy example for CB

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

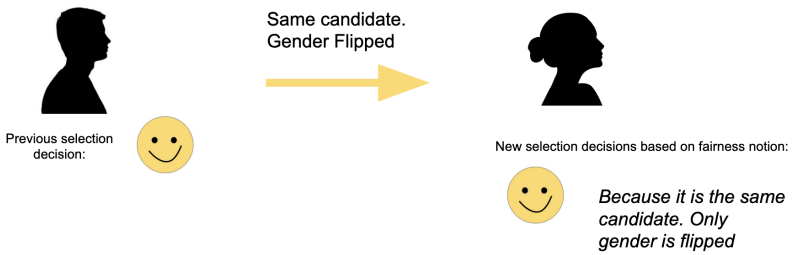
Example : There are 2 job applicants - one man and one woman with the same characteristics except gender. We either select both or reject both because they differ only on gender.



(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.



(c) Final design of the toy example for CF