Perceived Algorithmic Fairness using Organizational Justice Theory: An Empirical Case Study on Algorithmic Hiring

Guusje Juijn
guusjejuijn@hotmail.com
Utrecht University
Utrecht, the Netherlands

Niya Stoimenova
niya.stoimenova@deus.ai
DEUS
Amsterdam, the Netherlands

Joao Reis
joao.reis@deus.ai
DEUS
Porto, Portugal

Dong Nguyen
d.p.nguyen@uu.nl
Utrecht University
Utrecht, the Netherlands

ABSTRACT

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

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

algorithmic decision-making, organizational justice, perceived fairness, algorithmic hiring

ACM Reference Format:
Guusje Juijn, Niya Stoimenova, Joao Reis, and Dong Nguyen. 2023. Perceived Algorithmic Fairness using Organizational Justice Theory: An Empirical Case Study on Algorithmic Hiring. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3600211.3604677

1 INTRODUCTION

Artificial Intelligence systems are increasingly being used to inform and make important decisions about human lives across a wide range of high-impact domains, such as criminal law, medicine, finance, and employment [42]. While algorithmic decision-making has the potential to offer numerous promising advantages to society, such as increased efficiency and accuracy, it can also produce discriminatory or unfair outcomes [23, 32], as evidenced by several infamous cases such as COMPAS, the criminal risk assessment algorithm that was accused of being racially biased against black defendants [2], and Amazon's recruitment tool, which turned out to discriminate against female candidates [13]. Ensuring algorithmic fairness has therefore become a major area of interest within the field of artificial intelligence. This has led to the design of a whole landscape of fairness criteria and approaches to embed these into algorithms, as well as to the development of multiple bias mitigation algorithms, open-source libraries, and auditing toolkits to measure, visualize, and improve different fairness aspects [4, 6, 30, 39, 45].

However, there are still large gaps between fairness researchers and machine learning practitioners [36]. As it is impossible to mathematically satisfy all the proposed statistical fairness criteria at once since they are mutually incompatible [3, 9, 24], a universal consensus on how to ensure algorithmic fairness is lacking [44]. More knowledge about what criteria or metrics to use in what context is hence needed, which underscores the importance of approaching algorithmic fairness not only from a technical viewpoint. We need to understand what humans perceive as fair, to ensure that algorithmic decision-making systems are accepted, trusted, and supported by humans, since fairness is not purely an algorithmic concept, but a human construct [5, 7, 36, 42].

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08…$15.00
https://doi.org/10.1145/3600211.3604677

The literature on human perceptions of algorithmic fairness, however, frequently offers mixed or inconsistent results, highlighting the need for a more coherent approach to algorithmic fairness [12, 42]. Therefore, multiple studies have started to draw inspiration from organizational justice, which is concerned with fairness perceptions of decisions made about employees in organizational settings [14, 18, 19, 22, 27, 32]. Organizational justice literature divides fairness perceptions into three distinct but correlated components: distributive fairness, procedural fairness, and interactional fairness [17]. This categorization can therefore provide a solid foundation on how to systematically investigate algorithmic fairness perceptions.

However, most of the research into algorithmic fairness perceptions focuses merely on one of these three fairness components. In this work, we aim to investigate *the effect of integrating these components on algorithmic fairness perceptions.* Additionally, we investigate the link between mathematical algorithmic fairness and human perceptions of distributive fairness, by examining *whether participants have a preference for either demographic parity or equality of opportunity.* We focus on algorithmic hiring, a context that is easily comprehensible for a lay public. While this area has seen increased interest in the integration of AI-enabled software, it that has also witnessed raising concerns about the potential of AI to perpetuate or exacerbate existing biases [29, 35, 40]. As a result, it is classified as a high-risk area in the EU AI act [11]. Moreover, there is no universal agreement on how fairness should be formalized in algorithmic hiring: for instance, certain recruitment algorithms proactively aim to increase diversity when ranking job candidates, while others do not [16]. As research has demonstrated that fairness perceptions during a hiring process play a critical role in job satisfaction, performance, and the relationship between employers and employees, obtaining insights into the perceived fairness of algorithmic hiring is of particular importance [25].

Toward that end, we conduct an experiment with 225 predominantly White, native English Prolific crowdworkers from the UK, in which we examine fairness perceptions of several hypothetical recruitment algorithms. We study the following two research questions:

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

By grouping our participants based on the amount of information they receive about the recruitment algorithms, according to the fairness components described in organizational justice theory, we find that participants who only receive information about the distributive fairness of the algorithms have the lowest fairness perceptions. When participants receive information about both procedural and distributive algorithmic fairness, they perceive the algorithms as fairer: interestingly, we observe this effect both when the sensitive attribute gender is included as a main feature in the algorithms and when it is not.

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

By showing participants graphs that report the trade-offs between selection rate differences and false negative rate differences between two gender groups, we find a general preference for equality of opportunity over demographic parity. By qualitatively analyzing the rationales behind participants' fairness ratings, these findings are affirmed: a larger proportion of participants states to focus on qualification and false negatives, rather than selection rates. However, most participants specifically report taking into account the trade-offs between both these fairness criteria.

In sum, our study provides valuable insights into the relationship between algorithmic fairness and human perceptions of justice. Our experimental data and code can be found on our GitHub Repository: https://github.com/GuusjeJuijn/fairness-perceptions.

2 RELATED WORK

We start by taking a mathematical perspective on algorithmic fairness, by providing a concise overview of the most common criteria for algorithmic fairness and their associated trade-offs (§2.1). Subsequently, we adopt a human perspective, by describing the empirical literature on human algorithmic fairness perceptions and discussing the components of perceived fairness from organizational justice in an algorithmic context (§2.2).

2.1 Mathematical algorithmic fairness

Algorithmic fairness is a profoundly complex and many-faceted concept, which is reflected by the large landscape of criteria that try to grasp its meaning: with over 21 established mathematical formulas for fairness in binary classification problems, researchers have not yet come to a universal consensus on how to mathematically define what it means for a decision to be fair [8]. This section summarizes the fairness criteria that are most widely adopted and relevant to our study.

2.1.1 Group fairness. Group fairness criteria focus on treating persons that belong to a protected group, defined by a sensitive attribute such as gender or race, the same as persons that belong to any other group. To capture the different formulas belonging to this class, Barocas et al. [3] propose a taxonomy of statistical non-discrimination criteria consisting of three categories: independence, separation, and sufficiency. By depicting the sensitive attribute as S, the predicted outcome (the decision) as $\hat{Y}$, and the (true) outcome as Y, these three categories can be represented as follows:

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

Within independence, the most common fairness criterion is demographic parity, or disparate impact. A classifier satisfies this criterion when the percentage of favorable outcomes is equal for both the protected and unprotected group [31]. To adhere to demographic parity, the true outcome Y does not have to be known: for instance, in a hiring setting, a recruitment algorithm satisfies demographic parity between men and women when hiring an equal number of male and female candidates, regardless of their qualifications.

More complex definitions fall under separation and sufficiency. If the predicted outcome is conditionally independent of the sensitive attribute, given the true outcome, a classifier satisfies separation [3]. Two fairness criteria falling under this category are equality of opportunity, which requires the false negative rate to be equal for both groups, and predictive equality, which requires the false positive rate to be equal for both groups [9, 21].

Lastly, if the true outcome is conditionally independent of the sensitive attribute, given the predicted outcome, a classifier satisfies sufficiency. Sufficiency hence requires equal true outcomes over people that are given similar predictions. An example of a fairness criterion satisfying sufficiency, is calibration or test fairness [44].

2.1.2 Trade-offs between fairness definitions. Following the proliferation of research into mathematical criteria to define algorithmic fairness, researchers have started to investigate the mathematical relationships between these criteria. This has exposed an important issue: satisfying all fairness criteria simultaneously is impossible, as, under mild assumptions, any two out of the three aforementioned categories of group fairness are mutually exclusive [3, 9, 24]. Practitioners are therefore faced with the challenge of selecting among different fairness criteria and their associated trade-offs. However, which choice to make is a highly context-specific and difficult task, given the subtle differences between the different criteria, as well as other factors such as the availability of sensitive features, the level of understanding of the actual outcome label, and legal or organizational restrictions [36]. Multiple scholars, therefore, state that more emphasis on the social, human side of fairness is needed: in order to develop fair AI, it is essential to understand what humans perceive as fair and to acknowledge that fairness is not merely a technical construct [15, 36, 42].

2.2 Perceived algorithmic fairness

A growing body of literature applies organizational justice theory to the topic of perceived algorithmic fairness [5, 14, 18, 19, 22, 27, 32]. Organizational justice, like algorithmic decision-making, centers around the fairness of decisions made about others in a hierarchical environment. This similarity makes organizational justice a suitable source of inspiration for studying perceived algorithmic fairness [5]. Here, we discuss some of the related work on algorithmic fairness that focuses on one of the different components of perceived fairness described in organizational justice theory.

2.2.1 Distributive algorithmic fairness. Distributive fairness refers to the fairness of outcome distributions. It is based on norms for outcome allocation, such as equality (outcomes should be distributed equally amongst everyone) and equity (opportunities should be distributed equally based on everyone's circumstances) [10, 32, 42]. Robert et al. [37] note that distributive fairness is the most commonly discussed category within AI fairness literature. This finding could be attributed to the fact that many statistical fairness criteria emphasize distributive fairness, by focusing on how outcomes are divided across groups or individuals [32]. Dolata et al. [15] refer to this conclusion as the *distributiveness assumption*: the assumption that all fairness concerns can be represented as an outcome distribution problem. Most of the empirical work on the perceived fairness of algorithm outcomes focuses on basic fairness concepts, such as equality and equity [42]. However, only a handful of studies on distributive algorithmic fairness focus on the perceived fairness of particular mathematical fairness criteria specifically [22, 41].

Srivastava et al. [41] conduct an experiment to identify the mathematical fairness criterion that best captures crowdworkers' perceptions of fairness. By letting participants choose between a succession of model pairs, showing the predictions and true outcomes of a medical risk and criminal risk prediction algorithm, they find that participants prefer demographic parity over more complicated definitions, such as error parity and equal false positive rates. This finding suggests that humans exhibit a preference for fairness definitions that are more simplistic in nature.

However, Harrison et al. [22] draw different conclusions. They perform a between-subjects experiment in a bail decision-making context, in which they let participants judge the fairness of two models with pairwise fairness trade-offs. They identify two interesting fairness preferences: first, subjects favor equalizing the false positive rate over equalizing the accuracy across groups. Second, subjects also favor equalizing the false positive rate over equalizing the percentage of favorable outcomes (i.e., having demographic parity) across groups.

This latter result is in contrast with that of Srivastava et al., raising questions about the effect of different visualizations and ways of presenting information on participants' fairness perceptions.

2.2.2 Procedural algorithmic fairness. Unlike distributive fairness, procedural fairness focuses on the fairness of the decision-making process rather than the outcome. Morse et al. [32] investigate the procedural fairness of five popular mathematical fairness criteria along the six components of procedural fairness originally described by Leventhal [28]: consistency, bias suppression, representativeness, correctability, accuracy, and ethicality. By relating the fairness criteria to these different components, they provide directions for choosing the right criterion per situation and provide a fundament for better understanding and assessing the procedural fairness of these fairness metrics: they, for example, reason that equality of opportunity and equalized odds are criteria with a high level of procedural fairness [32].

Grgic-Hlaca et al. [18] take a different approach to investigate procedural algorithmic fairness: they seek to identify feature properties that influence the perceived fairness of using certain features as input for an algorithmic decision-making model. By investigating participants' assessments of different feature properties, they find that participants consider a feature's perceived relevance and reliability most important. As these feature properties are unrelated to discrimination, Grgíc-Hlaca et al. conclude that procedural unfairness concerns reach far beyond discrimination only and that therefore, other feature properties should also be taken into account when assessing algorithmic fairness.

Other authors explore the perceived procedural fairness of including certain features in an algorithm. Pierson [34], for example, finds that men are more likely to include gender as an attribute in an education recommendation algorithm, compared to women. Grgić-Hlača et al. [20] moreover find that men perceive the inclusion of race as a feature as more fair compared to women.

2.2.3 Interactional algorithmic fairness. Lastly, interactional fairness refers to providing sufficient information and giving truthful explanations about decision procedures. It is concerned with presenting people with adequate information about the process of how a decision is reached and is therefore closely related to procedural fairness[1] [5, 10]. In an organizational justice setting, an example of interactional fairness is providing employees explanations for layoff decisions: it has been shown that if employees receive honest, thorough, and accurate explanations when being fired, they perceive these decisions as significantly fairer [27].

Multiple studies investigate the effect of explanations for decisions on perceived algorithmic fairness. For example, by performing a user study in a criminal risk setting, Dodge et al. [14] find that feature importance-based explanations and demographic-based explanations increase participants' algorithmic fairness perceptions. In an online user study in a medical decision-making context, Angerschmid et al. [1] also find a positive effect of feature importance-based explanations on perceived algorithmic fairness. These insights will be leveraged in **RQ1** of our study.

3 METHODOLOGY

Our methodology is two-folded. We first created machine learning models that adhered to different fairness criteria (§3.1). We then conducted an online user study in which participants judged the fairness of these models (§3.2).

3.1 Model development

We first trained machine learning models on the Utrecht Fairness Recruitment dataset[2]. As this data set was specifically designed to mimic realistic recruiting data and to demonstrate fairness issues, and did not contain any missing values or ambiguous features, we considered it an appropriate data set for the purposes of our user study. The data set contained information about the recruitment decisions of four hypothetical companies. We split the data from one company into a training set (750 instances) and a testing set (250 instances). Using Scikit-learn [33], we trained three logistic regression models, using default parameters, to predict whether an individual in the data set was hired by the company or not. We trained one original, raw model, one model mitigated for demographic parity, and one model mitigated for equality of opportunity. Bias mitigation was applied using the ThresholdOptimizer algorithm[3] from Microsoft FairLearn [6]. This postprocessing algorithm, introduced by Hardt et al. [21], adjusts a learned classifier by applying group-specific thresholds, to satisfy a specified fairness constraint.

Postprocessing for demographic parity and equality of opportunity specifically was done for several reasons. First of all, multiple studies suggest that both of these criteria are appropriate for algorithmic hiring, the context we focus on in our empirical study [16, 26, 32, 35]. Mitigating for demographic parity, moreover, allowed for further investigation of the results of Srivastava et al. [41], who found that lay people tend to have a preference for this criterion in different contexts. Besides, as demographic parity is often used in practice and relatively easy to understand, we considered this to be a suitable criterion for this study [38]. Since, according to Morse et al. [32], equality of opportunity scores high on procedural fairness, we considered this a second suitable criterion. The accuracies and fairness metrics of all three classifiers are reported in Table 1. Although the mitigated models did not perfectly meet the proposed criteria, postprocessing substantially decreased the differences in either selection rates or false negative rates between groups.

3.2 Empirical study

To assess our research questions, we performed an online experiment on the crowdsourcing platform Prolific Academic using Qualtrics survey software. The survey was distributed at the end of January 2023. Here, we outline our study design, survey structure, and participant demographics.

3.2.1 Study Design. Participants' fairness perceptions of several hypothetical recruitment algorithms were assessed using a direct measure based on Harrison et al. [22], asking *"Do you think this algorithm is fair?"*. To ensure that every participant had a similar definition in mind, we provided them with a fairness definition by Mehrabi et al. [31]: *"Fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits"*. Participants were asked to provide a judgment on a 7-point Likert scale, ranging from 1 ("not at all fair") to 7 ("completely fair"). Additionally, at the end of the survey, participants were asked to elaborate on the motivations behind their ratings through an open-ended query, asking *"In the previous questions, which factors did you consider most important in determining whether an algorithm was fair or unfair?"*. This was done to qualitatively investigate the rationales behind the respondents' fairness perceptions.

Each participant was presented with five different graphs representing algorithms, of which the selection rates and false negative rates were based on the logistic regression models described in §3.1. In these graphs, the selection rates were defined as the proportion of hired candidates, while the false negative rates were defined as the proportion of qualified candidates who were not hired. We explicitly opted to describe the figures in this way, as we anticipated that the terms 'selection rate' and 'false negative rate' would not be easily comprehensible to participants without machine learning knowledge.

One of these five algorithms represented the original, unmitigated model. Two of these algorithms represented demographic parity: one perfectly following the criterion and one representing the mitigated model. Two of these algorithms represented equality of opportunity: again, one perfectly following the criterion and one representing the mitigated model.

Participants were divided into three groups. The amount of information participants received about these algorithms differed per group, based on the fairness components described in organizational justice theory. We considered procedural and interactional fairness together, due to the strong connection and overlap between these two components.

Group 1: distributive fairness. The first group only received information about the distributive fairness of the algorithms. This was visualized as a graph representing the algorithm outcomes,

[1] In our empirical study, we, therefore, choose to consider procedural and interactional fairness together.
[2] https://www.kaggle.com/datasets/ictinstitute/utrecht-fairness-recruitment-dataset
[3] https://fairlearn.org/v0.8/user_guide/mitigation.html

Table 1: Fairness metrics of the original model, the demographic parity-mitigated model and the equality of opportunity-mitigated model

Model	Gender	Accuracy	Selection Rate	False Negative Rate
Original	Female	0.918	0.123	0.333
	Male	0.847	0.468	0.158
	Difference	*0.071*	*0.335*	*0.175*
Demographic parity-mitigated	Female	0.839	0.197	0.133
	Male	0.742	0.250	0.509
	Difference	*0.151*	*0.035*	*0.376*
Equality of opportunity-mitigated	Female	0.926	0.164	0.133
	Male	0.847	0.468	0.158
	Difference	*0.079*	*0.304*	*0.025*

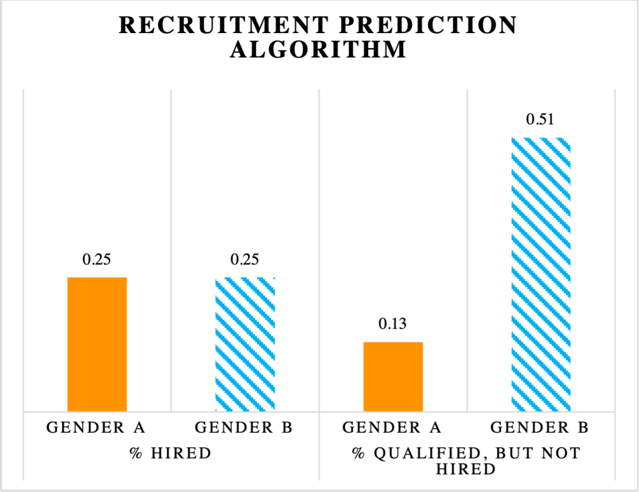

Figure 1: Example outcome graph, representing distributive fairness, showed to each participant. On the left, the selection rates are shown. On the right, the false negative rates are shown. This algorithm adheres to demographic parity but not to equality of opportunity.

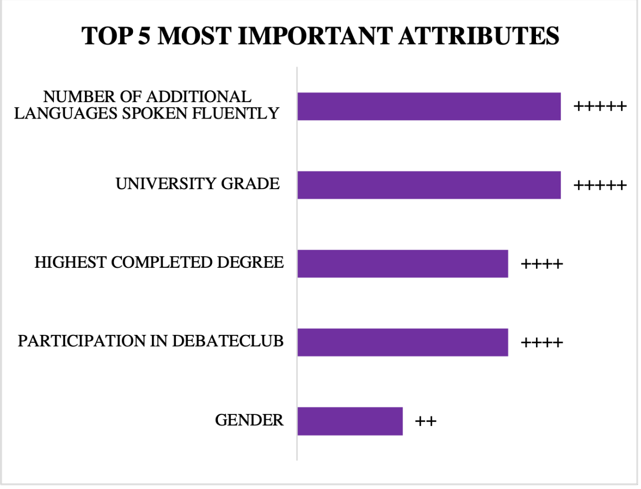

Figure 2: Feature importance graph shown to group 2, representing procedural fairness. The graph shown to group 3 was the same, except for the sensitive attribute 'gender' being changed for the non-sensitive attribute 'exact study'.

showing a pairwise trade-off between the selection rates and false negative rates between two gender groups. Instead of only showing one aspect of algorithmic fairness, by, for example, only showing the difference in false negative rates between groups, we chose to represent a more realistic real-world scenario by showing the trade-offs between different fairness criteria. By doing so, we drew inspiration from the work of Harrison et al. [22]. We explicitly chose to rename the two gender groups into *Gender A* and *Gender B*, to limit the effect of implicit biases regarding gender roles. An example of a graph representing distributive fairness is shown in Figure 1.[4]

Group 2: distributive and procedural fairness, with sensitive attribute. The second group not only received information about the distributive fairness of the algorithms, but also about the procedural fairness of the algorithms. Like Grgic-Hlaca et al. [18], we considered the features used by the algorithm as an important aspect of procedural fairness. Therefore, we visualized procedural fairness as a feature importance explanation. Like Dodge et al. [14], we presented the feature coefficients of the logistic regression models as strings of '+'s representing the relative importance of each feature. To limit the amount of information, we only showed the top five most influential features. For each of the algorithms, the feature importance graph stayed the same, as postprocessing does not change the model coefficients. Figure 2 displays the feature importance graph shown to the participants of group 2.

[4] We first piloted these graphs amongst colleagues, to make sure they were clear enough to interpret.

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

3.2.2 Survey Structure. After signing a consent form, participants were shown an introductory text. The purpose of this text was to introduce the topic of algorithmic fairness, clarify the task, present the context, and demonstrate a sample graph to ensure that the participants could properly interpret the visual representations. Each participant was then randomly assigned to one of the three groups. The participants were divided evenly across the groups to ensure that each group had an equal number of participants. Within each group, every participant was asked to rate the fairness of five different recruitment algorithms: one representing the original, unmitigated model, two adhering to demographic parity, and two adhering to equality of opportunity. These algorithms were presented in a randomized order to limit order effects. After these five questions, participants were asked to write down which factors they considered most important in their fairness analysis. The survey ended with demographic questions and a message thanking the participants for their time and giving them a completion code to register their submission in Prolific. Figure 3 shows an overview of the experimental flow.

3.2.3 Participants. Participants were pre-screened on having obtained at least a high school diploma, having English as a first language, and residing in the UK. We rewarded them with £10,84 per hour, conforming to the minimum wage in the UK. On average, the survey took 4.2 minutes to complete. By manually checking the response times, data from participants that took less than 2 minutes to complete the survey were deleted to ensure the quality of answers. In total, data from 225 participants were used. Table 2 summarizes our participants' demographics[5].

Table 2: Participants' demographics

		% (n=225)
Gender	Female	50%
	Male	50%
	Other	<1%
Age	18–30	33%
	30–45	35%
	45–60	22%
	60+	10%
Race/ethnicity	White	92%
	Asian	4%
	Mixed	3%
	Black	1%
Education	High school diploma	54%
	Technical/community college	40%
	Undergraduate degree	5%
	Graduate degree	<1%
	Doctorate degree (PhD/other)	<1%

4 RESULTS

4.1 Quantitative Analysis

4.1.1 ***RQ1.*** First, we considered the effect of the type of information given about the algorithms on participants' fairness perceptions. For each of the three groups, we computed the average fairness perceptions of the original algorithm, the algorithms representing demographic parity, and the algorithms representing equality of opportunity. As shown in Figure 4, participants who received information about both the distributive and procedural fairness of the algorithms (groups 2 and 3) consistently perceived the algorithms as fairer compared to participants who only received information about the distributive fairness of the algorithms (group 1). We observed this effect in both groups 2 and 3, although fairness perceptions were generally higher in group 3, in which the sensitive attribute gender was not included as a main attribute in the feature importance graph.

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

4.1.2 ***RQ2.*** Next, we investigated whether participants preferred either demographic parity or equality of opportunity. For each of the three groups, we computed the average fairness perceptions of the algorithms representing demographic parity and the average fairness perceptions of the algorithms representing equality of opportunity. Figure 5 shows that across all three groups, participants tended to have a preference for the algorithms representing equality of opportunity. A Wilcoxon-Signed Rank test (a non-parametric variant of the paired t-test) indicated that in groups 2 and 3, the average perceived fairness scores for the algorithms representing equality of opportunity were significantly higher than the average perceived fairness scores for the algorithms representing demographic parity *($W = 601.5, p = 0.013$* and *$W = 636.5, p = 0.016$* respectively). However, in group 1, these differences were not statistically significant *($W = 777.0, p = 0.541$).*

4.1.3 ***Gender differences.*** Additionally, we examined potential differences in average scores among male and female participants. Of all three groups, for each algorithm, we compared the average scores between men and women, using a Mann-Whitney U-test (a non-parametric variant of the independent t-test). However, the results of these tests did not reveal any significant differences.

[5] Age and race were automatically collected by Prolific. Our survey additionally asked for gender and the highest level of education obtained.

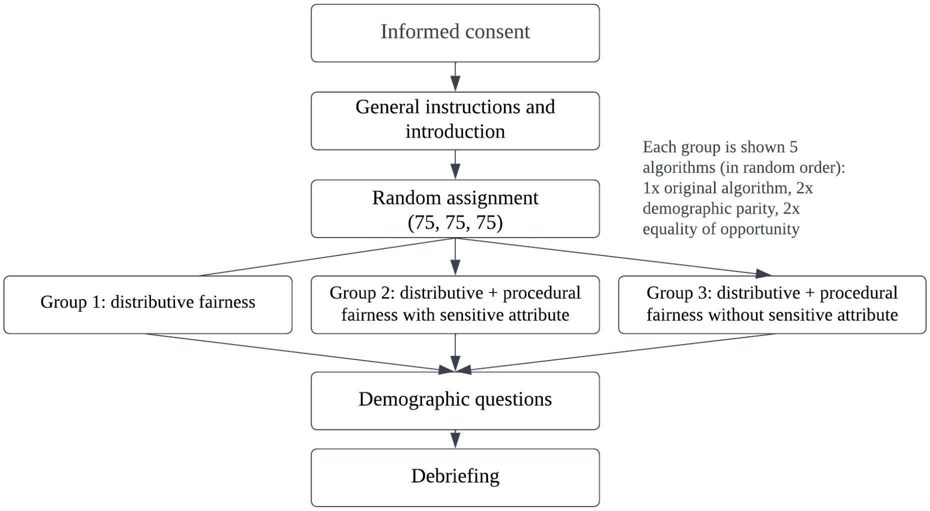
Informed consent
General instructions and introduction
Random assignment (75, 75, 75)
Each group is shown 5 algorithms (in random order): 1x original algorithm, 2x demographic parity, 2x equality of opportunity
Group 1: distributive fairness
Group 2: distributive + procedural fairness with sensitive attribute
Group 3: distributive + procedural fairness without sensitive attribute
Demographic questions
Debriefing

Figure 3: Experimental Flow

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

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

4.2 Qualitative Analysis

To gain additional insights into the findings of our quantitative analysis, we qualitatively analyzed participants' rationales behind their fairness ratings by openly coding their responses to the open-ended question of which factors they considered most important in determining the fairness of the algorithms. Although each participant provided an explanation, we encountered a variety of response lengths: responses varied in length between 1 word and 59 words, with a mean of 12 words and a median of 9 words. By first identifying first-order codes out of these responses and grouping these into second-order codes, we systematically classified the responses. Figure 6 gives an overview of these categories and provides, per category, an indicative quote. Two annotators independently reviewed the responses. In 80% of the cases, they initially agreed. The remaining 20% of responses were assigned a final classification after a discussion between the annotators. For 9% of the responses, no clear category was identified (e.g.: *"If it looked fair or not", "All combined"*). In 4 responses, multiple categories were mentioned. In these cases, our approach was to classify the response based on the category mentioned first.

We now discuss some of the responses falling under the two second-order codes we identified: distributive fairness and procedural fairness.

4.2.1 *Distributive fairness.*

While we encountered a variety of answers, the biggest proportion of explanations (n=164, 73%) could be attributed to the outcome of the algorithms, relating to the concept of distributive fairness. This was as expected, as only two out of three groups received a feature importance graph, and all three groups received information about distributive fairness. However, interestingly, we observed that across all three groups, the majority of participants focused on distributive fairness rather than procedural fairness (82% of all answers in group 1, 61% of all answers in group 2, and 76% of all answers in group 3).

More specifically, across all three groups, we found that most participants (n=68) emphasized the importance of considering the trade-offs between the different fairness criteria shown in the graphs. For example, P45 (group 1), stated: *"I mainly looked at the proportions between genders of those qualified but not hired in comparison to the genders when hired"*. The second most frequently mentioned category pertained to the concept of equal opportunity:

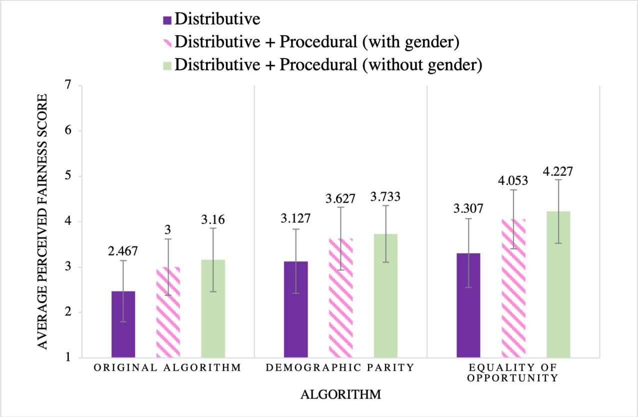

Figure 4: Average perceived fairness scores, on a 7-point Likert scale, of each of the three groups. Error bars indicate standard deviations. Bar graphs show that the group that only received information about the distributive fairness of the algorithms rated each of the three algorithms lower than the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

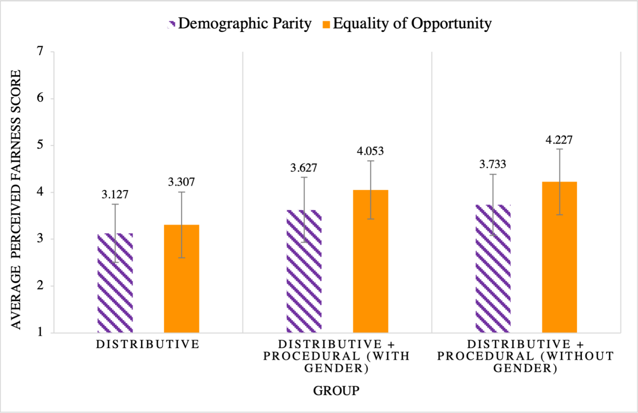

Figure 5: Average perceived fairness scores, on a 7-point Likert scale, of the algorithms adhering to demographic parity and equality of opportunity. Error bars indicate standard deviations. Bar graphs show that across all three groups, algorithms adhering to equality of opportunity were rated higher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

a notable proportion of participants (n=53) mainly focused on false negative rates and the qualifications of candidates. This finding suggests a preference for fairness criteria that consider the actual outcome. For example, P18 (group 1) answered: *"The percentage that was qualified but not hired was the most important factor for me"*. Nevertheless, there was also a considerable number of participants (n=43) that primarily considered the selection rates of both groups, e.g.: *"Whether the hired % of candidates were as equal as possible"* (P38, group 2). However, across all three groups, this category, associated with demographic parity, was mentioned less frequently than the category relating to equal opportunity.

4.2.2 Procedural fairness. 18% of answers (n=41) could be attributed to the decision-making process, and therefore, to the concept of procedural fairness (7% of all answers in group 1, 27% of all answers in group 2, and 21% of all answers in group 3).

The majority of these responses (n=34) were related to the features used by the algorithms and their relative importance. For example, in group 2, in which gender was included as a main attribute in the feature importance graph, we encountered 11 answers that explicitly criticized its usage, e.g.: *"I marked them all low as I don't see why gender would be an important factor "* (P68, group 2). Other participants mainly focused on the importance or combination of the different attributes, e.g., *"The 5 main attributes were the main thing I considered"* (P52, group 2).

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

5 DISCUSSION

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

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

Our findings underscore the potential consequences of adopting the *distributiveness assumption* as described by Dolata et al. [15], as we show that solely representing algorithmic fairness as an outcome distribution issue can lead to lower perceptions of fairness. Our results suggest that providing more information about the workings of an algorithm can enhance fairness perceptions. This is consistent with the results of Dodge et al. [14] and Angerschmid et al. [1], who found that feature importance-based explanations have a positive impact on algorithmic fairness perceptions.

Furthermore, our work provides empirical insights into how mathematical fairness criteria are related to human algorithmic fairness perceptions. By measuring and comparing participants' fairness perceptions of recruitment algorithms adhering to two

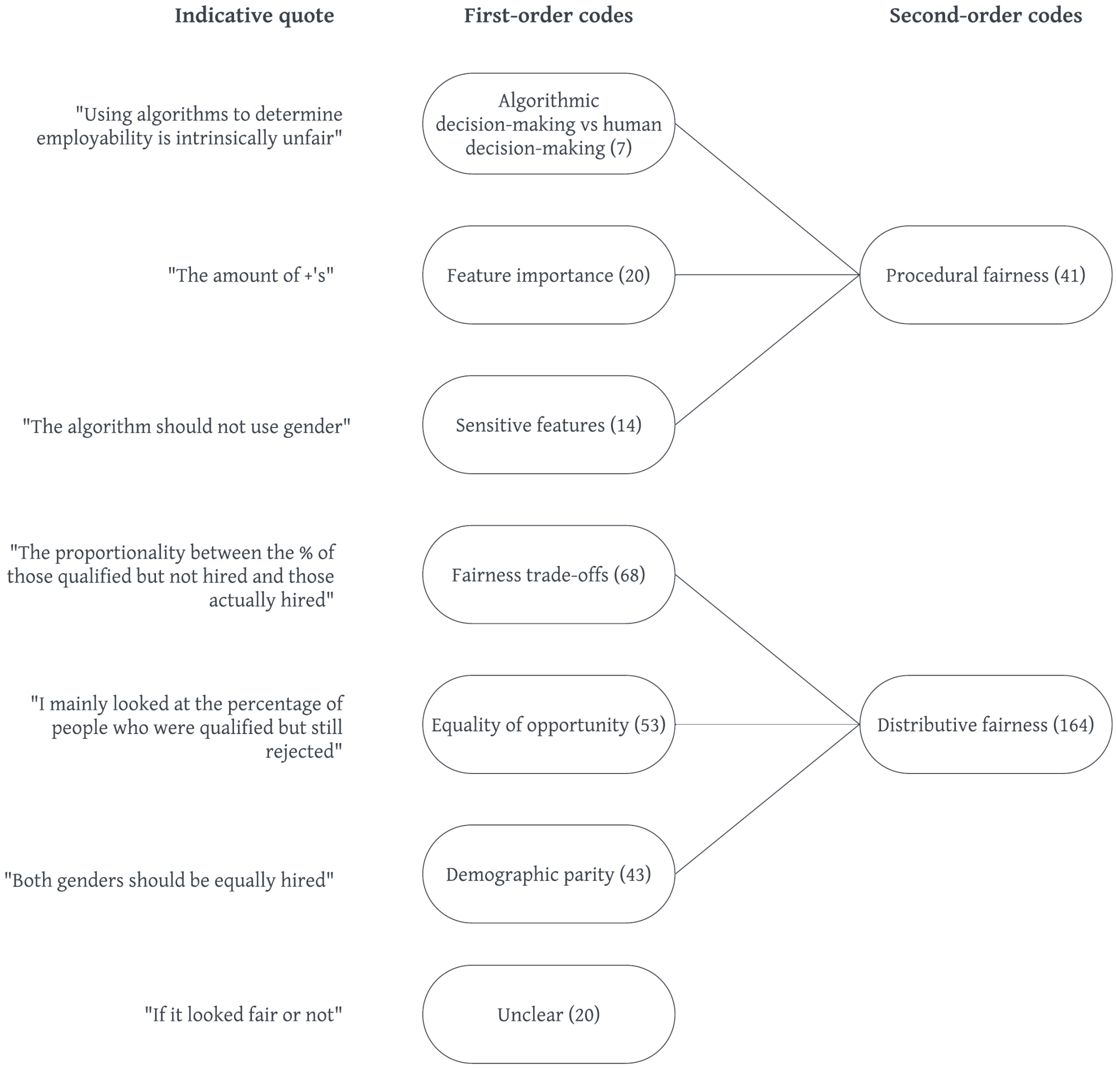
Indicative quote
First-order codes
Second-order codes
"Using algorithms to determine employability is intrinsically unfair"
Algorithmic decision-making vs human decision-making (7)
"The amount of +'s"
Feature importance (20)
Procedural fairness (41)
"The algorithm should not use gender"
Sensitive features (14)
"The proportionality between the % of those qualified but not hired and those actually hired"
Fairness trade-offs (68)
"I mainly looked at the percentage of people who were qualified but still rejected"
Equality of opportunity (53)
Distributive fairness (164)
"Both genders should be equally hired"
Demographic parity (43)
"If it looked fair or not"
Unclear (20)

Figure 6: Indicative quotes, first-order codes, and second-order codes for the open-ended question: *"Which factors did you consider most important in determining whether a model was fair or unfair?"*

different algorithmic fairness criteria, we find **a significant preference for equality of opportunity over demographic parity**, when given information about both the distributive and procedural fairness of the algorithms. These findings are affirmed in our qualitative analysis, in which we note that a larger proportion of participants assigns greater importance to false negative rates when forming their fairness judgments, as opposed to (equal) selection rates among genders.

Our results are in contrast with the preference for demographic parity found by Srivastava et al. [41]. As they focus on a medical risk prediction and criminal risk prediction setting, rather than hiring, these varying contexts could be a possible reason behind these contrasting findings. For instance, decision-making in medical and criminal risk settings may involve higher stakes compared to hiring. Moreover, these settings may not capture the imagination as much as hiring does, possibly leading to different fairness judgements. It is, however, also plausible that these contrasting results can be explained by the varying methods of visualizing fairness issues. Where Srivastava et al. [41] represent their algorithms by showing the individual outcomes of ten decision subjects, we report the trade-offs between two fairness criteria. Moreover, while all participants in the study of Srivastava et al. [41] are solely provided with information about the algorithmic outcomes, relating to the concept of distributive fairness, two-thirds of our participants also receive information about the procedural fairness of the algorithms.

Another potential explanation for our findings could be associated with the participants' levels of comprehension of the fairness criteria. In a study into lay people's understanding of mathematical fairness criteria, interestingly, Saha et al. [38] find that participants' comprehension of equality of opportunity is lower compared to their comprehension of demographic parity. Additionally, they observe that participants who score higher on comprehension tend to have lower fairness perceptions. In line with this reasoning, a possible explanation for our findings is that our participants had a better understanding of the algorithms adhering to demographic parity compared to the algorithms adhering to equality of opportunity. This could have resulted in assigning a lower score to the algorithms adhering to demographic parity.

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

A second limitation pertains to our approach to describing false negative algorithmic predictions in terms of qualifications. We used synthetic data in our experiments. However, in real-world hiring scenarios, determining whether candidates are 'qualified' is a subjective decision, susceptible to different types of biases. It is important to acknowledge that a real-world hiring scenario encompasses a much greater level of complexity, in which qualifications may never be assessed with complete certainty.

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

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

Future Directions. Our results emphasize that understanding algorithmic fairness perceptions requires careful consideration of both visualization and contextual factors. Suggestions for future work, therefore, include:

- Exploring the effect of presenting various visualizations, and offering additional context about the decision-making process, on participants' algorithmic fairness evaluations. Van Berkel et al. [43], for example, take a useful start in this direction, by evaluating the effect of scatterplot and text-based visualizations of algorithmic outcomes on fairness perceptions.
- Assessing participants' algorithmic fairness perceptions using implicit measures, rather than explicitly asking whether they think an algorithm is fair. Implementing such a design could potentially reduce the influence of cognitive biases and response biases, such as social desirability bias.
- Investigating participants' preferred mathematical fairness criteria in multiple contexts, besides algorithmic hiring. For instance, a future study could categorize various contexts based on the risk-oriented approach of the AI act, which categorizes AI systems into 4 levels: unacceptable, high, minimal, or low risk [11]. Such a study could then examine whether participants' preferences for certain fairness criteria in different contexts vary based on these different levels of risk.
- Studying whether participants' fairness perceptions are affected by receiving additional information about an algorithm, by conducting a within-subjects study, as opposed to a between-subjects study. For example, one approach could involve presenting participants with information about the distributive fairness of an algorithm, followed by information about its procedural fairness. By asking for their fairness perceptions at these two points in time, it could be investigated whether providing information about procedural fairness *alters* fairness perceptions.

6 CONCLUSION

In this study, we approach the topic of perceived algorithmic fairness through the lens of organizational justice theory, using algorithmic hiring as a case study. Our key finding is that providing information about the procedural fairness of an algorithm increases fairness perceptions, even when the process can be considered unfair. We moreover find a preference for equality of opportunity over demographic parity, when given information about the distributive and procedural fairness of an algorithm. Our results highlight the interplay between the different components of fairness in organizational justice theory, and the relationship between mathematical algorithmic fairness and perceived algorithmic fairness. By performing an empirical study amongst crowdworkers, we add to the growing body of literature on public perceptions of algorithmic fairness and provide important directions for future research.

ACKNOWLEDGMENTS

We want to thank the anonymous reviewers for their useful feedback. Moreover, we thank Maartje de Graaf, Rosanna Nagtegaal, Sieuwert van Otterloo, Goya van Boven, Michael Pieke, Daan van der Weijden, and Tim Koornstra for their advice on this project.

REFERENCES

[1] Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. 2022. Fairness and explanation in AI-informed decision making. *Machine Learning and Knowledge Extraction* 4, 2 (2022), 556–579.

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

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[5] 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. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[6] 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. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[7] Alycia N Carey and Xintao Wu. 2022. The statistical fairness field guide: perspectives from social and formal sciences. *AI and Ethics* (2022), 1–23.

[8] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (2022), 1–21.

[9] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[10] Jason A Colquitt. 2001. On the dimensionality of organizational justice: a construct validation of a measure. *Journal of applied psychology* 86, 3 (2001), 386.

[11] European Commission. 2023. *Regulatory framework proposal on Artificial Intelligence.* Retrieved March 13, 2023 from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[12] Sophia T Dasch, Vincent Rice, Venkat R Lakshminarayanan, Taiwo A Togun, C Malik Boykin, and Sarah M Brown. 2020. Opportunities for a More Interdisciplinary Approach to Perceptions of Fairness in Machine Learning. In *NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA)*.

[13] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[14] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE 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*. 275–285.

[15] Mateusz Dolata, Stefan Feuerriegel, and Gerhard Schwabe. 2022. A sociotechnical view of algorithmic fairness. *Information Systems Journal* 32, 4 (2022), 754–818.

[16] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2221–2231.

[17] Jerald Greenberg. 1987. A taxonomy of organizational justice theories. *Academy of Management review* 12, 1 (1987), 9–22.

[18] 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*. 903–912.

[19] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[20] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M. Redmiles. 2022. Dimensions of diversity in human perceptions of algorithmic fairness. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*. Article 21, 12 pages.

[21] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[22] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 392–402.

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

[24] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67)*. 43:1–43:23.

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

[26] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. 2022. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research* 297, 3 (2022), 1083–1094.

[27] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[28] Gerald S Leventhal. 1980. What should be done with equity theory? In *Social exchange*. Springer, 27–55.

[29] 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.

[30] Trisha Mahoney, Kush Varshney, and Michael Hind. 2020. *AI Fairness.* O'Reilly Media, Incorporated.

[31] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[32] Lily Morse, Mike Horia M Teodorescu, Yazeed Awwad, and Gerald C Kane. 2021. Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics* (2021), 1–13.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[34] Emma Pierson. 2017. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124* (2017).

[35] 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*. 469–481.

[36] Brianna Richardson and Juan E Gilbert. 2021. A framework for fairness: a systematic review of existing fair AI solutions. *arXiv preprint arXiv:2112.05700* (2021).

[37] 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 (2020), 545–575.

[38] Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. 2020. Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning*. PMLR, 8377–8387.

[39] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[40] Candice Schumann, Jeffrey Foster, Nicholas Mattei, and John Dickerson. 2020. We need fairness and explainability in algorithmic hiring. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[41] 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*. 2459–2468.

[42] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (2022), 20539517221115189.

[43] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[44] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.

[45] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.