10-9-2023

Human perceptions of fairness: a survey experiment

Julian Sengewald
Technische Universität Dortmund, julian.sengewald@tu-dortmund.de

Anissa Schlichter
Technische Universität Dortmund, anissa.schlichter@tu-dortmund.de

Markus Siepermann
Technische Hochschule Mittelhessen, markus.siepermann@mni.thm.de

Richard Lackes
Technische Universität Dortmund, richard.lackes@tu-dortmund.de

Recommended Citation

Sengewald, Julian; Schlichter, Anissa; Siepermann, Markus; and Lackes, Richard, "Human perceptions of fairness: a survey experiment" (2023). *Wirtschaftsinformatik 2023 Proceedings*. 72.
https://aisel.aisnet.org/wi2023/72

Human Perceptions of Fairness: a Survey Experiment

Research Paper

Julian Sengewald[1], Anissa Schlichter[1], Markus Siepermann[2], and Richard Lackes[1]

[1] Technical University Dortmund, Business Informatics, Dortmund, Germany
{julian.sengewald,anissa.schlichter,richard.lackes}@tu-dortmund
[2] University of Applied Sciences, Mathematics, Natural Sciences and Computer Science, Giessen, Germany
markus.siepermann@mni.thm.de

Abstract. Algorithmic decision-making (ADM) through automation has benefits but must be implemented responsibly. Several mathematical definitions of fair outcomes exist, but it remains unclear how these align with human perceptions of fairness. We conducted a survey experiment (N=258) examining common machine-learning definitions of fairness (demographic parity, equal opportunity, and equalized odds) in the context of algorithmic job interview invitations. We find that humans perceive the simple fairness definition of demographic parity as less fair than a more complex one that considers whether the invitees were eligible.

Keywords: perceived fairness, survey experiment, algorithmic decision-making.

1 Introduction

Nowadays, algorithmic decision-making (ADM) is being used prominently in business operations with high visibility and high liability. A prime example is recruitment, where ADM is used for candidate selection and resume screening (Linkedin Talent Solutions 2018, 40–50) – where nondiscriminatory behavior is a concern.

Organizations prefer ADM because they deem it a more objective hiring process (Linkedin Talent Solutions 2018, 40; van den Broek et al. 2019, 3–4). However, ADM can perpetuate bias and discrimination in recruitment because it relies on historical data that may not represent a diverse workforce (Dastin 2018; Köchling et al. 2021; Köchling and Wehner 2020; Mehrabi et al. 2021). As a result, ADM performance can be impaired and unfair for different demographic groups. Despite the existence of various computational approaches to address this problem (Friedler et al. 2019), these approaches currently lack a behavioral empirical foundation.

While humans have an innate sense of what constitutes fair treatment from an early age (Sloane et al. 2012), incorporating this sense into an algorithm is challenging. Previous studies (Cheng et al. 2021; Harrison et al. 2020; Srivastava et al. 2019; Wang et al. 2020) have used an empirical approach to identify mathematical definitions of fairness that correlate with *perceived fairness*. By identifying the moral standard that people use to evaluate the distributive outcome of ADM, a machine learning (ML) model can then be "tuned" accordingly, e.g., to choose more equitably to improve its perceived fairness. However, perceptions of fairness are highly context-dependent and subject to varying moral standards. But the existing literature has primarily examined the perceived fairness of ADM in criminal justice and health care while recruitment is a distinct domain. First, a large applicant pool may make it necessary to reject even qualified candidates, a trade-off that organizations must balance. Furthermore, ADM errors in high-stakes domains such as bail, pretrial release, or cancer detection can have irrevocable and serious consequences. In recruitment, errors have less of an impact because candidates who are rejected can re-apply to other companies. Consequently, findings from the areas examined in the prior literature, where decision-making is typically one-off, may not be directly applicable to ADM recruitment. Therefore, we pose the following research question:

RQ: *Which mathematical definition of fairness in ML most closely correlates to people's empirical perception of fairness in ADM-based recruitment?*

By answering this research question, we contribute to the research on how to leverage the potential of ADM in recruitment (Langer et al. 2019; Lee 2018; Suen et al. 2019; van Esch et al. 2019). For this , we explore how mathematical fairness definitions for ML (Harrison et al. 2020; Srivastava et al. 2019; Wang et al. 2020) are perceived in a novel application domain and address to the recent demand for more responsible and ethical ADM (van der Aalst 2017). This paper is organized as follows: In Section 2, we describe how ADM can be used in recruitment and provide the technical background for the mathematical definition of fairness. The experimental setup and design are outlined in Section 3.1. Section 3.2 describes how perceived fairness was measured, and Section 3.3 describes the empirical strategy. The results are presented in Section 4 and discussed in Section 5.

2 Background

2.1 Fairness in ADM and ML

Advances in ML have enabled widespread ADM adoption. Yet, ML models use probabilistic rules, which can lead to biased-erroneous decisions that treat individuals unfairly (Feuerriegel et al. 2020). Ethical issues arise when these errors disparately affect certain demographic groups. ADM's connection to fairness lies in the benefit allocated by its decision-making, e.g., when making hiring decisions (benefit = job) or deciding about university admissions (benefit = study place); all areas where unfair treatment would be problematic. If an ML classifier assigns a benefit (e.g., a hiring decision), we denote this as $\hat{Y} = 1$ and otherwise as $\hat{Y} = 0$, if no benefit is assigned. The classifier is learned from historical data, where $Y = 1$ indicates that an individual with characteristics X was eligible for the benefit and $Y = 0$ if the individual was ineligible. The decision-making of the ML classifier and eligibility is naturally linked to a confusion matrix; therefore, some authors also use the expression *predictive fairness* when discussing the matter, e.g., (Haas 2019)

There are three ways to ensure that the output of an ML classifier is fair:

- **Pre-processing:** alters data before training, e.g., by reweighting (Kamiran and Calders 2012). This works with any ML classifier.
- **At training time/in-processing:** a method that achieves high utility by simultaneously optimizing the classifier and fairness, e.g., using regularization (Kamishima et al. 2011). The downside is the need for specialized software.
- **Post-processing:** adjusting the predicted label of a classifier after prediction, e.g. using a linear program (Hardt et al. 2016).

For each method, the user must specify a mathematical fairness definition, a *fairness metric*, which can be classified in group- and individual-based metrics (Hutchinson and Mitchell 2019, 52). Group-based metrics are prevalent in software tools (e.g., IBM's AI Fairness 360, Bellamy et al. 2019, Microsoft's Fairlearn, Bird et al. 2020, and Google's What-If tool, Wexler et al. 2020). These metrics calculate the distribution of benefits between groups $g_1, \dots, g_i, \dots, g_G$. Gender, age, or ethnicity are examples of attributes used to form groups. If the benefit is allocated equally across the groups, according to a fairness metric, we say that *parity* is achieved for this metric.

Demographic Parity (DP) asserts that an ML classifier is fair if it gives the benefit to all demographic groups at the same rate τ^{DP}:

$$P\left(\hat{Y} = 1 \middle| Group = g_i\right) = \tau^{DP} \; \forall \; g_i, i \in 1, \dots, G$$

A DP classifier does not systematically disadvantage, or advantage individuals based on what group they are in. Unlike a quota system, which is common in the workplace (for example, requiring at least 20% of an executive board to be women), DP requires equal allocation rates for all groups. Although DP ensures the independence of the group and benefit allocation, it does not consider eligibility. The allocation rate among eligible individuals is measured by true-positive-rate ($TPR \coloneqq P(\hat{Y} = 1|, Y = 1)$), among non-eligible by false-positive-rate, ($FPR \coloneqq P(\hat{Y} = 1|, Y = 0)$). Allocation rates can be conditioned on eligibility and group to define yet another notion of fairness.

Equal Opportunity/True-Positive Parity (TPP) requires an equal benefit distribution over demographic groups conditioned on eligibility (Hardt et al. 2016). That means that equal opportunity holds if the group-specific TPR has the same value τ^{OP} across groups. True positive parity (TPP) holds when TPR is the same for all groups:

$$P\left(\hat{Y} = 1 \middle| Y = 1,\ g_i\right) = \tau^{OP}, \forall \; g_i \; with \; i \in \{1, \dots, G\}$$

Equalized Odds (TPP + FPP) entails, in addition to TPP, also equalized False-Positive Rate (FPR), i.e. False Positive Parity (FPP) holds (Hardt et al. 2016):

$$P\left(\hat{Y} = 1 \middle| Y = 1, g_i\right) = \tau_1^{OD} \text{ and } P\left(\hat{Y} = 1 \middle| Y = 0, g_i\right) = \tau_2^{OD}, \forall \; g_i \; i \in \{1, \dots, G\}$$

Other parity metrics are Accuracy Parity (ACCP), $P\left(\hat{Y} = \mathrm{Y} \middle| Group = g_i\right) = \tau^{acc}$, and False Negative Parity (FNP), $P\left(\hat{Y} = 0 \middle| Y = 1, Group = g_i\right) = \tau^{\mathrm{FNP}}$, $\forall \; g_i \; i \in \{1, \dots, G\}$.

2.2 Review of the literature

The literature has increasingly focused on the issue of the perceived fairness of ADM. While perceived fairness is a broad concept (Dolata et al. 2022), the fourfold model (procedural, interpersonal, informational and distributive) is an empirically supported conceptualization of it (Colquitt and Shaw 2005). Studies have examined procedural features of ADM in terms of perceived fairness, such as how it compares to human decision-making depending on the task being automated (Lee 2018) and the role of data sources (Albach and Wright, James, R. 2021; Grgic-Hlaca et al. 2018). Also, informational aspects such as providing explanations and transparency contribute to perceived fairness (Binns et al. 2018; Dodge et al. 2019; Wang et al. 2020). Providing explanations to improve perceived fairness, while not significant in the case of being self-affected (Wang et al. 2020), can increase the perceived fairness of the overall ADM if designed as global explanations, whereas local explanations tailored to a specific outcome are more appropriate to justify a particular decision (Dodge et al. 2019). Another line of research focuses on distributive fairness, concerning how well ADM is perceived to treat those affected by its decision (Hannan et al. 2021; Harrison et al. 2020; Saxena et al. 2020; Srivastava et al. 2019; Wang et al. 2020). This study contributes to the distributive fairness and group-based metrics literature. Distributive fairness is measured by fairness metrics, which can be categorized into individual-based fairness metrics (Hannan et al. 2021; Saxena et al. 2020), and group-based metrics (Harrison et al. 2020; Srivastava et al. 2019; Wang et al. 2020).

A summary of the research on group-based metrics is shown in Table 1. The first study used two settings and an adaptive design (see Table 1), where the fairness metrics were generated to match the participant's preferred fairness metric. Participants preferred DP in both scenarios (Srivastava et al. 2019). Harrison et al. (2020) investigated how people perceive the fairness of different ways to make an ML model fair, by group metrics (ACCP and FNP[1]) or by the raw model score (i.e., no fairness definition). The study was conducted in the context of deciding whether to grant bail. The experiment was designed as a binary comparison between two models that had one or both properties equalized. Their data indicate a small, but statistically significant, preference for FNP rates between demographic groups (Harrison et al. 2020). Wang et. al. (2020) considered how (un-) equal error rates affect the fairness perception of ML algorithms. As equal error rates imply equal accuracy across groups, we remapped the results of Wang et. al. to ACCP in Table 1. Furthermore, the results showed that the perception of being self-disadvantaged by a biased algorithm was independent of the participant demographics. Also, participants perceived the algorithm as unfair if they received an unfavorable outcome, even though the algorithm was unbiased. Lastly, recent research

[1] In general, in binary ML classification the positive class can be arbitrary assigned to one of the classes. In the study of Harrison et al. (2020) a FP was labelled as “A defendant that is mistakenly denied bail […]” (p. 4). In this case the decision linked to the positive class is “deny bail” and the beneficial outcome for the defendant would be “grant bail” is the negative class here. Most other studies labelled the beneficial outcome as the positive class (Srivastava et al. 2019). To be consistent across the literature we define in the positive class as the beneficial outcome and decided to remap FP to FN accordingly when necessary.

has chosen a purely qualitative approach (Cheng et al. 2021). When comparing DP and EOP, their finding suggests a preference for EOP (Cheng et al. 2021). Most studies involving perceived fairness of group-based metrics deal with relatively high stake settings (e.g., crime and health) (Cheng et al. 2021; Harrison et al. 2020; Srivastava et al. 2019), and just a few studies deal with low-stakes settings (Wang et al. 2020). Participants are mostly required to select among two ML models satisfying either one/some (conflicting) fairness definitions. Their role was mainly a passive rater not self-affected by the ADM (Harrison et al. 2020; Srivastava et al. 2019). Typically, the fairness definitions were preset (Harrison et al. 2020, 396; Wang et al. 2020, 5), but an adaptive design has also been chosen (Srivastava et al. 2019).

In contrast, we use a new domain in a low stake setting, in which using ADM is well accepted (Langer et al. 2019). Unlike prior studies, we also utilize anonymous groupings to promote neutral responses. In addition, the set of metrics has not be shown to participants in this combination.

Table 1. Literature Review

Literature	Setting	Approach	Predictive fairness					
			ACCP	FPP	FNP	DP	FPP & TPP	Sample
(Srivastava et al. 2019)	Crime, Health	• Options: Model trade-offs • Measurement: Binary choice • Perspective: Passive rater • Groups: Gender, race • Visualization: Confusion Matrix	○	○	○	●		Online AMT[1] N=200
(Harrison et al. 2020)	Bail	• Options: Model trade-offs • Measurement: 5-point Likert scale • Perspective: Passive rater • Groups: Race • Visualization: Diagram/Percentage	○		●			Online AMT N=502
(Wang et al. 2020)	Training	• Options: Overall rating • Measurement: 7-point Likert scale • Perspective: Personal affected • Groups: Gender, age, race • Visualization: Diagram/Percentage	●					Online AMT N=590
(Cheng et al. 2021)	Social care	• Qualitative Study • Perspective: Active • Groups: Gender, age, race • Visualization: Multidimensional				○	●	Experts N=12
This study	Recruitment	• Options: Overall rating • Measurement: 5-point Likert scale • Perspective: Passive rater • Groups: anonymous • Visualization: Confusion matrix		x		x	x	Online volunteers N=258

○ not-most-preferred, ● most preferred. Abbreviations:

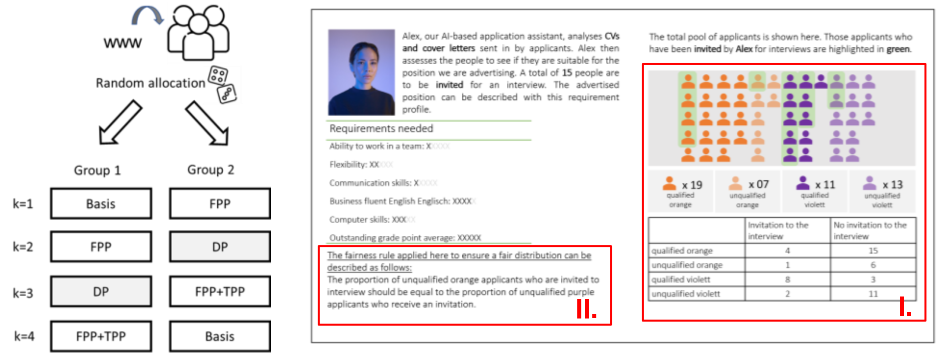
ACCP Accuracy Parity. FNP False Negative Parity.

FPP False Positive Parity. DP Demographic Parity.

3 Methods

3.1 Experiment

Experimental Approach. We use an online survey experiment as our study method which is – according to the literature – the established method (see Figure 1). We conducted a factorial survey, also known as “vignette experiments” (Atzmüller and Steiner 2010), to determine what the most preferred fairness definition is. In a factorial survey, participants are shown a full description of the scenario and are therefore well suited for studying social norms (Alves and Rossi 1978).



Alex, our AI-based application assistant, analyses **CVs and cover letters** sent in by applicants. Alex then assesses the people to see if they are suitable for the position we are advertising. A total of **15** people are to be **invited** for an interview. The advertised position can be described with this requirement profile.

Requirements needed

Ability to work in a team: X

Flexibility: XX

Communication skills: X

Business fluent English Englisch: XXXX

Computer skills: XXX

Outstanding grade point average: XXXXX

The fairness rule applied here to ensure a fair distribution can be described as follows:
The proportion of unqualified orange applicants who are invited to interview should be equal to the proportion of unqualified purple applicants who receive an invitation.

The total pool of applicants is shown here. Those applicants who have been **invited** by **Alex** for interviews are highlighted in **green**.

	Invitation to the interview	No invitation to the interview
qualified orange	4	15
unqualified orange	1	6
qualified violett	8	3
unqualified violett	2	11

(a) Experimental layout (b) Exemplary vignette maximized for DP

Figure 1. Experimental structure and vignette design.

Four vignettes were presented to each participant in a within-subject design (see Figure 1 panel (a)). Each vignette showed the outcome of an ADM using a confusion matrix. We generated each confusion matrix by a constraint optimization problem, maximizing either the equalized false positive rate, equalized chances, or demographic parity while holding overall accuracy and the number of invitations constant. By fixing accuracy, we ensure that participants only evaluate distributive changes not the overall algorithm performance (Srivastava et al. 2019). We also supplied a fairness-unoptimized baseline case. Participants scored three fairness criteria and the baseline scenario. The vignettes were built on each other, so their sequencing was not randomized. FPP was displayed first because it is nested within FPP + TPP. The case DP, unrelated to the nested scenarios, was added between them to hide the nesting. This allowed participants to easily transition between vignettes.

Setting. As the usage domain of ADM, we used candidate prescreening for an advertised job offer. To make the setting realistic, the number of job candidates invited to a personal interview is limited, and therefore only the most qualified candidates should be invited. This use of ADM is one where decision-making is fully automated (Langer et al. 2019). The scenario was chosen for two reasons: first, the experiment's participants would presumably be familiar with the setting of a job interview. This increases the internal validity of the given answers. Furthermore, candidate prescreening for interview invites provides study participants with a realistic and compelling scenario.

Vignette Design. Figure 1 panel (b) shows an exemplary vignette. The vignette begins with a general description: Using ADM for candidate prescreening along with an imaginary job profile. Whether a job candidate is invited to an interview depends on whether the ADM system deems them qualified. The vignette also shows two groups and their corresponding ADM outcomes. Selected job candidates are highlighted in green, while their qualification level is depicted by intense or non-intense shading. Additionally, the algorithmic outcome is also depicted using two group-specific confusion matrices. The confusion matrix was varied as follows (Figure 1, panel (b), I.):

Table 2. Variation of Invitation (Rows: vignettes. Columns: Confusion matrices)

	TP		TN		FP		FN	
	orange	violet	orange	violet	orange	violet	orange	violet
Basis	9	3	4	13	3	0	10	8
FPP	4	8	6	11	1	2	15	3
DP	5	7	4	13	3	0	14	4
FPP+TPP	8	4	6	11	1	2	11	7

We also give a short explanation of the related fairness definition (Figure 1, panel (b), II.), to rule out that some definitions are harder for participants to understand (Saha et al. 2020). These descriptions are (translated from German):

- *Basis*. "The ratio of invited qualified applicants to the not-invited qualified applicants should be significantly greater than the ratio of unqualified applicants who receive an invitation to unqualified applicants who are not invited to interview".
- *FFP*. "The proportion of unqualified orange applicants who are invited to interview should be equal to the proportion of invited unqualified purple applicants".
- *FPP+TPP*. FPP and "The proportion of not-invited qualified orange applicants should equal the proportion of not-invited qualified purple applicants".
- *DP*. "The proportion of orange applicants who receive an invitation to interview should be equal to the proportion of purple applicants who receive an invitation".

We depicted both groups solely by color (orange/purple) to maintain anonymity and prevent response biases. In-group favoritism (e.g., male participants give a higher rating to outcomes favoring male depicted candidates) (Harrison et al. 2020, 401) or support for affirmative action (Harrison et al. 2020, 398; Saxena et al. 2020, 5–6), are known predictors of response bias. Blinding the group mitigates the response bias.

3.2 Measurement

We use "overall fairness" to measure the perceived fairness among study participants. Perceived overall fairness can be measured by the individual's experience of fairness and the overall entity perception (Colquitt and Shaw 2005). We use a reduced version of an existing scale that combines both types of fairness ratings (Colquitt and Shaw 2005), with a total of three measurement items. For each item, participants reported on a six-point Likert scale ranging from 1 (‘*Completely disagree*’) to 6 (‘*Completely agree*’). Table 3 shows the items and participants' average ratings for each item.

The overall fairness perception scale is reflective and gives the measurement model: $y_j = \lambda_j \eta + \varepsilon_j$. Each reflective measurement y_j obtained by item j measures with a correlation λ_j the underlying construct η plus an item-specific error term ε_j. To evaluate the reflective model, two factors were examined: 1) the internal consistency reliability, and 2) convergent validity (Hair et al. 2021). First, using the criteria of composite reliability ρ_c and Cronbach’s α, we assess the internal consistency. The composite reliability ρ_c is reported to be 0.89, indicating that there is a high level of internal consistency (Hair et al. 2021). We also calculated Cronbach’s α, the average inter-item correlation, which has a value of 0.81, indicating that there is a high level of consistency between the items (Robinson 2018). Next, we assess the convergent validity. For this validation, we use the average variance extracted (AVE). The AVE is 0.72, which is higher than the required minimum of 0.5 (Hair et al. 2016)

Table 3. Overall fairness, (Colquitt and Shaw 2005), α =0.81, ρ_c =0.89, AVE=0.72

Item	**Text**	**Mean**	**Std. dev.**
1	I feel that the way the decision was made was fair.	3.48	1.07
2	Overall, the decision-making process treats every candidate fair.	3.46	1.01
3	Most candidates would say that they were treated fair in the decision-making process.	3.34	1.06

3.3 Empirical strategy

For our analysis, we used an *ordinal mixed-effects* regression model. An ordinal model is a reliable method for estimating the effect of categorical, ordinal, and metric variables on an ordinal outcome (McCullagh 1980). In contrast to a linear model, the ordinal model is more robust against false alarms and false omissions when applied to ordinal data (Liddell and Kruschke 2018). We use a *multinomial logistic* model using a logistic link function, e.g., (Christensen 2018). In our context, the multinomial logistic model assumes a latent fairness perception score y^* for which only categorical realization on the Likert scale is observable. This means participants choose a particular value c on the Likert scale if their latent fairness perception resides in the interval (θ_{c-1}, θ_c). The observed distribution of the fairness ratings, conditioned on the vector of observable characteristics, can therefore be expressed as $P(y_{ij,k} = c|x) = P(\theta_{c-1} < y^*_{ij,k} \leq \theta_c|x)$. For answering our research question, we explore how the different trade-offs of optimizing the fairness definitions are perceived.

Owing to the peculiar experimental structure, repeated measurements, we construct an appropriate empirical model. There are repeated measurements for the items and participants. Participants rated a total of four different vignettes. The response pattern of the participant may depend not only on the vignette and observable demographics but also on many unobserved participant characteristics. These unobserved characteristics are modeled by a participant-specific error term $\dot{u}_i$. We also included a random effect for each of the four repeated measures for the item j and those repeated measures share an error term from the reflective measurement model $y_j = \lambda_j \eta + \varepsilon_j$. For repeated measurements, a mixed-effects model is better suited, ensuring more efficient estimates and sound statistical inference (Atzmüller and Steiner 2010). We denote individuals by i, with $i \in 1, \dots, N$, the item j, with $j \in 1,2,3$, and trial k, with $k \in 1,2,3,4$. The full mixed-effects model is given as follows:

$$y_{ij,k=}\beta_o + \beta_1\, case + \beta_2\, age + \beta_3\, sex + \beta_4\, education + \dot{u}_i + \varepsilon_j + u_{ij,k}$$

Besides the mixed effects model, we include a linear model in the results, because it is widely used, and to illustrate the robustness of our findings. The linear model is estimated by OLS. Because of repeated measurements, we report clustered standard errors using the sandwich package in R (Zeileis et al. 2020). Standard errors are clustered by the participant, and we use a degree-of-freedom correction of $n/(n-k)$ for the covariance matrix (MacKinnon and White 1985; Zeileis et al. 2020, 7–9). The dependent variable is based on averaging the reflective items $y_{i,k}^{OLS} = \frac{1}{3}\sum_{j=1}^{3} y_{ij,k}$.

4 Results

4.1 Sample

Study participants were voluntarily recruited on social media (Facebook, and WhatsApp) and via e-mail lists. To encourage participation, each participant was offered an incentive in the form of a charitable donation. In total, 523 answered the online questionnaire. 258 participants completed the fairness ratings. The analysis sample was these 258 participants. The sample consists of 58% female, 36% male, and 6% of an undisclosed gender. The age distribution is young, with 64% of the participants being 23-30 years old and 16% being 22 years of age or younger. About 6% did not disclose their gender. The sample is highly educated, with 78% having finished or currently pursuing a university education, which is a higher level of education than the German average (DeStatis 2020). We also asked how frequently study participants had experienced discrimination in work-related decisions such as hiring or job promotion. Female study participants reported being discriminated against in 46% whereas male ones reported 42%. For the entire sample, 42% of the participants had experienced discrimination, and 11% had experienced ‘often’ discrimination ‘often’ in the workplace. Therefore, the sample shows awareness of the study topic.

4.2 Fairness preferences

We now turn to the empirical results. Table 4 reports the results of four regression models. In the first model, we include the main explanatory variable, “optimizing for a mathematical fairness definition”, and no further controls. According to the computed coefficients, the TPP fairness criterion stands out as the most well-received option, as it is widely seen as the only agreement that meets the criteria of high perceived fairness. In contrast, other options were found to fall short of these requirements.

Next, we included control variables for demographic characteristics. Overall, the estimated coefficients remain robust in terms of their relative ranking and numerically. The combined effect of the definition of fairness of equalized odds (FPP and TPP) is positive (beta = 0.9, se = 0.133, p < 0.05), which we obtained from conducting an ordinal regression containing an appropriately specified dummy coefficient.

Table 4. Regression results

	Ordinal (Coefficients are log odds)			**OLS**
Variables\Model	(1)	(2)	(3)	(4)
Fairness definitions (dummy coded)				
True positive parity (TPP)	1.218***	1.218***	1.219***	0.660***
	(0.097)	(0.097)	(0.097)	(0.094)
Demographic parity (DP)	-0.162*	-0.161*	-0.161*	-0.067
	(0.093)	(0.093)	(0.093)	(0.079)
False positive parity (FPP)	-0.301***	-0.301***	-0.302***	-0.140**
	(0.093)	(0.093)	(0.093)	(0.067)
Controls				
Age				
<=30		-0.343	-0.343	-0.119
		(0.329)	(0.329)	(0.185)
> 30		-0.221	-0.222	-0.119
		(0.386)	(0.386)	(0.185)
Missing		-0.195	-0.195	-0.130
		(0.601)	(0.601)	(0.315)
Gender				
Female		0.390*	0.390*	0.184*
		(0.207)	(0.207)	(0.104)
Missing		1.515**	1.516***	0.747**
		(0.649)	(0.648)	(0.297)
Education				
Bachelor or in-education		0.037	0.037	0.045
		(0.268)	(0.268)	(0.125)
Master and above		0.012	0.012	0.011
		(0.324)	(0.324)	(0.160)
Missing		-1.125	-1.125	-0.557*
		(0.811)	(0.811)	(0.304)
Items	RE	RE	FE	Averaged
Participants	RE	RE	RE	RE
N	3096	3096	3096	1032
Nakagawa's conditional R2	0.465	0. 463	0.463	-
R2	-	-	-	0.06
Significance levels: *** p≤0.01, **p≤0.05, *p≤0.1				

In general, all demographic control variables (age, gender, and education) are not statistically significant. Unlike the other models, Model 3 controls for participants' perceived fairness with fixed effects. We found that the previous results hold regardless of whether the reflective measure is included as a fixed or random effect term. The fixed effects for the measurement items are not significantly different from those for the intercept, which includes one of the measurement items. The non-significant result can be interpreted as indicating that the items behave similarly, demonstrating the good properties of our reflective scale for measuring perceived fairness. Finally, we also include a linear model that specifies the Likert scale as the metric variable. The estimated coefficients can now be interpreted directly on a scale from 1 to 7. We again find that the TPP is perceived as more fair than the baseline case. The signs of demographic parity and FPP are negative, indicating that they are perceived as less fair. The effect size is small, and the coefficient is not statistically significant. We also examined possible heterogeneity and tested for interaction effects between demographics and fairness definitions but found no significant effect. In sum, to ensure fairness in simple decisions, TPP criterion is preferable over others.

5 Discussion

The analysis suggests that participants perceive the definitions of equalized odds (FPP and TPP) to be fairer than the remaining definitions. Similar results to ours were obtained in the experiment by Cheng et al. (2021) with qualitative methods. Our approach supports this finding using a quantitative approach and in a recruitment context. Furthermore, we find that DP is the second most preferred fairness metric. These results are not in line with those of Srivastava et al. (2019), who found that DP was considered the fairest. We also find that FPP is the least preferred fairness definition. This finding can be interpreted as follows: participants understood the experiment well because the unqualified would have been invited at the expense of the qualified. It also means that participants considered the business utility of the selection procedure, although we hold accuracy constant. We include FPP in the researched set of fairness metrics because FPP can reduce workplace discrimination by detecting phenomena such as in-group favoritism and nepotism (offering resources preferably to family and friends). We find mixed significance in the results for demographics; similar to previous work (Wang et al. 2020). In summary, the results show that perceived fairness can be increased by choosing an appropriate fairness definition to make ML fair; enterprises can increase the perceived fairness of their ADM-based recruitment processes to meet organizational objectives like organizational attractiveness, job satisfaction, commitment, and turnover intentions (Ambrose and Schminke 2009; McCarthy et al. 2017). The fairness definition DP is the second fairest, and therefore may be regarded as a good and easy-to-realize proxy for fairness in the recruitment process. However, the introduction of a quota system into recruitment processes based solely on this definition appears not recommendable. In addition, the DP fairness definition may not accurately capture the fairness perception of participants. It is also possible that participants were predisposed to engage in the DP fairness definition because of their familiarity with government employment advertisements, where the promotion of equality (e.g., between genders) is commonplace. Altogether, what is noticeable is that even for the most regarded fairness definition, equalized odds, the overall fairness perception of the ADM outcome is low. This finding somewhat implies that the set of fairness metrics is not exhaustive and future research is needed. Lastly, we did not find significant age which is line (Wang et al. 2020). The same holds for education which stands in contrast to previous studies (Wang et al. 2020). This could be due to the sample composition.

As with any study, there are limitations. First, the study participants did not rate the actual use of ADM in recruitment. However, due to the realistic scenario, the overall setting is very believable to participants. Furthermore, we used an online volunteer sample, and a different population may provide different ratings. However, students have a higher thematic involvement in the scenario because they will likely encounter ADM-based recruitment. We studied a small set of fairness definitions, and the inclusion of a new fairness metric in the experiment may give another ranking. Overall, the sample is characterized by people of young age and higher education and a large proportion of women. Therefore, the findings would need to be validated with a broader sample. To increase the generalizability of our findings, we propose that future research tests our findings in a different cultural setting because social norms could influence what is perceived as fair (Awad et al. 2018). Also, the organization should consider that solely optimizing for a fairness definition is not enough, because the impacted parties should also be informed about how fairness was ensured, as in our experiment. It would be interesting to study how the design of the information impacts the perception of fairness. Organizations should also consider procedural factors when designing their ADM-based recruitment (Langer et al. 2019; Lee 2018; Ochmann et al. 2020).

6 Conclusion

In particular, ADM has been accepted in less complex decision-making scenarios, such as the prescreening of candidates (Langer et al. 2019). However, while ADM raises ethical concerns, it also allows us to think more deeply about previously implicit processes. Human decisions have more variability and subjectivity than ADM. Algorithmic processes, on the other hand, can be used to make algorithms behave according to societal norms when combined with crowdsourced perceptions of fairness. Therefore, this research can help companies make ADM-based hiring perceived as fair. In line with regulations such as the European Union's proposed AI Act, companies can mitigate the societal impact of AI while increasing efficiency through the use of automation. However, we also believe that excessive crowdsourcing should be scrutinized. Crowdsourcing fairness perceptions are consistent with characterizing existing societal norms (descriptive ethics), but embedding crowdsourced norms into algorithms also has normative implications that require caution.

References

Albach, Michele/Wright, James, R. (2021). The role of accuracy in algorithmic process fairness across multiple domains. Proceedings of the 22nd ACM Conference on Economics and Computation, 29–49. https://doi.org/10.1145/3465456.3467620.

Alves, Wayne M./Rossi, Peter H. (1978). Who Should Get What? Fairness Judgments of the Distribution of Earnings. American Journal of Sociology 84 (3), 541–564. https://doi.org/10.1086/226826.

Ambrose, Maureen L./Schminke, Marshall (2009). The role of overall justice judgments in organizational justice research: a test of mediation. The Journal of applied psychology 94 (2), 491–500. https://doi.org/10.1037/a0013203.

Atzmüller, Christiane/Steiner, Peter M. (2010). Experimental Vignette Studies in Survey Research. Methodology 6 (3), 128–138. https://doi.org/10.1027/1614-2241/a000014.

Awad, Edmond/Dsouza, Sohan/Kim, Richard/Schulz, Jonathan/Henrich, Joseph/Shariff, Azim/Bonnefon, Jean-François/Rahwan, Iyad (2018). The Moral Machine experiment. Nature 563 (7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Bellamy, Rachel K. E./Dey, Kuntal/Hind, Michael/Hoffman, Samuel C./Houde, Stephanie/Kannan, Kalapriya/Lohia, Pranay/Martino, Jacquelyn/Mehta, Sameep/Mojsilović, Aleksandra (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development 63 (4/5), 4: 1-4: 15.

Binns, Reuben/van Kleek, Max/Veale, Michael/Lyngs, Ulrik/Zhao, Jun/Shadbolt, Nigel (2018). 'It's Reducing a Human Being to a Percentage'. In: Regan Mandryk/Mark Hancock (Eds.). Engage with CHI. CHI 2018 : proceedings of the 2018 CHI Conference on Human Factors in Computing Systems : April 21 -26, 2018, Montréal, QC, Canada, CHI '18: CHI Conference on Human Factors in Computing Systems, Montreal QC Canada, 21 04 2018 26 04 2018. New York, New York, The Association for Computing Machinery, 1–14.

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

Cheng, Hao-Fei/Stapleton, Logan/Wang, Ruiqi/Bullock, Paige/Chouldechova, Alexandra/Wu, Zhiwei Steven Steven/Zhu, Haiyi (2021). Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In: Yoshifumi Kitamura/Aaron Quigley/Katherine Isbister et al. (Eds.). Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21: CHI Conference on Human Factors in Computing Systems, Yokohama Japan, 08 05 2021 13 05 2021. New York, NY, USA, ACM, 1–17.

Christensen, Rune Haubo B. (2018). Cumulative link models for ordinal regression with the R package ordinal. Submitted in J. Stat. Software 35.

Colquitt, Jason/Shaw, John (2005). How Should Organizational Justice Be Measured? In: Jerald Greenberg/Jason Colquitt (Eds.). Handbook of organizational justice. Mahwah, N.J, Lawrence Erlbaum Associates, 113–152.

Dastin, Jeffrey (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Available online at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (accessed 6/5/2022).

DeStatis (Ed.) (2020). Bildungsstand der Bevölkerung - Ergebnisse des Mikrozensus 2019 - Ausgabe 2020.

Dodge, Jonathan/Liao, Q. Vera/Zhang, Yunfeng/Bellamy, Rachel K. E./Dugan, Casey (2019). Explaining models. In: Wai-Tat Fu/Shimei Pan/Oliver Brdiczka et al. (Eds.). IUI '19, IUI '19: 24th

International Conference on Intelligent User Interfaces, Marina del Ray California, 17 03 2019 20 03 2019. New York (NY), ACM, 275–285.

Dolata, Mateusz/Feuerriegel, Stefan/Schwabe, Gerhard (2022). A sociotechnical view of algorithmic fairness. Information Systems Journal 32 (4), 754–818. https://doi.org/10.1111/isj.12370.

Feuerriegel, Stefan/Dolata, Mateusz/Schwabe, Gerhard (2020). Fair AI. Business & Information Systems Engineering 62 (4), 379–384. https://doi.org/10.1007/s12599-020-00650-3.

Friedler, Sorelle A./Scheidegger, Carlos/Venkatasubramanian, Suresh/Choudhary, Sonam/Hamilton, Evan P./Roth, Derek (2019). A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, GA, USA, Association for Computing Machinery, 329–338.

Grgic-Hlaca, Nina/Redmiles, Elissa M./Gummadi, Krishna P./Weller, Adrian (2018). Human Perceptions of Fairness in Algorithmic Decision Making. In: Pierre-Antoine Champin/Fabien Gandon/Mounia Lalmas et al. (Eds.). Proceedings of the 2018 World Wide Web Conference, the 2018 World Wide Web Conference, Lyon, France, 23.04.2018 - 27.04.2018. Republic and Canton of Geneva, International World Wide Web Conferences Steering Committee, 903–912.

Haas, Christian (2019). The price of fairness. A framework to explore trade-offs in algorithmic fairness. ICIS 2019 Proceedings 19. Available online at https://aisel.aisnet.org/icis2019/data_science/data_science/19/ (accessed 9/21/2020).

Hair, Joseph F./Hult, G. Tomas M./Ringle, Christian M./Sarstedt, Marko (2016). A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). 2nd ed. Thousand Oaks, SAGE Publications, Incorporated.

Hair, Joseph F./Hult, G. Tomas M./Ringle, Christian M./Sarstedt, Marko/Danks, Nicholas P./Ray, Soumya (2021). Evaluation of Reflective Measurement Models. In: Joseph F. Hair/Tomas M. Hult/Christian M. Ringle et al. (Eds.). Partial least squares structural equation modeling (PLS-SEM) using R. A workbook. Cham, Springer, 75–90.

Hannan, Jacqueline/Winnie Chen, Huei-Yen/Joseph, Kenneth (2021). Who Gets What, According to Whom? An Analysis of Fairness Perceptions in Service Allocation. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. Virtual Event, USA, Association for Computing Machinery, 555–565.

Hardt, Moritz/Price, Eric/Srebro, Nathan (2016). Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain, Curran Associates Inc, 3323–3331.

Harrison, Galen/Hanson, Julia/Jacinto, Christine/Ramirez, Julio/Ur, Blase (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. New York, NY, USA, Association for Computing Machinery, 392–402.

Hutchinson, Ben/Mitchell, Margaret (2019). 50 Years of Test (Un)fairness: Lessons for Machine Learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, GA, USA, Association for Computing Machinery, 49–58.

Kamiran, Faisal/Calders, Toon (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33 (1), 1–33. https://doi.org/10.1007/s10115-011-0463-8.

Kamishima, Toshihiro/Akaho, Shotaro/Sakuma, Jun (2011). Fairness-aware Learning through Regularization Approach. 2011 IEEE 11th International Conference on Data Mining Workshops, 643–650.

Köchling, Alina/Riazy, Shirin/Wehner, Marius Claus/Simbeck, Katharina (2021). Highly accurate, but still discriminatory. Business & Information Systems Engineering 63 (1), 39–54. https://doi.org/10.1007/s12599-020-00673-w.

Köchling, Alina/Wehner, Marius Claus (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), 795–848. https://doi.org/10.1007/s40685-020-00134-w.

Langer, Markus/König, Cornelius J./Papathanasiou, Maria (2019). Highly automated job interviews: Acceptance under the influence of stakes. International Journal of Selection and Assessment 27 (3), 217–234. https://doi.org/10.1111/ijsa.12246.

Lee, Min Kyung (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society 5 (1), 1-16.

Liddell, Torrin M./Kruschke, John K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? Journal of Experimental Social Psychology 79, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009.

Linkedin Talent Solutions (2018). Global recruiting trends 2018. The 4 ideas changing how you hire. Available online at https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/linkedin-global-recruiting-trends-2018-en-us.pdf (accessed 6/5/2022).

MacKinnon, James G./White, Halbert (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics 29 (3), 305–325. https://doi.org/10.1016/0304-4076(85)90158-7.

McCarthy, Julie M./Bauer, Talya N./Truxillo, Donald M./Anderson, Neil R./Costa, Ana Cristina/Ahmed, Sara M. (2017). Applicant Perspectives During Selection: A Review Addressing “So What?,” “What’s New?,” and “Where to Next?”. Journal of Management 43 (6), 1693–1725. https://doi.org/10.1177/0149206316681846.

McCullagh, Peter (1980). Regression Models for Ordinal Data. Journal of the Royal Statistical Society: Series B (Methodological) 42 (2), 109–127. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x.

Mehrabi, Ninareh/Morstatter, Fred/Saxena, Nripsuta/Lerman, Kristina/Galstyan, Aram (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys 54 (6), Article 115. https://doi.org/10.1145/3457607.

Ochmann, Jessica/Michels, Leonard/Zilker, Sandra/Tiefenbeck, Verena/Laumer, Sven (2020). The influence of algorithm aversion and anthropomorphic agent design on the acceptance of AI-based job recommendations. ICIS 2020 Proceedings. Available online at https://aisel.aisnet.org/icis2020/is_workplace_fow/is_workplace_fow/4.

Robinson, Mark A. (2018). Using multi-item psychometric scales for research and practice in human resource management. Human Resource Management 57 (3), 739–750. https://doi.org/10.1002/hrm.21852.

Saha, Debjani/Schumann, Candice/Mcelfresh, Duncan/Dickerson, John/Mazurek, Michelle/Tschantz, Michael (2020). Measuring non-expert comprehension of machine learning fairness metrics. In: International Conference on Machine Learning. PMLR, 8377–8387.

Saxena, Nripsuta Ani/Huang, Karen/DeFilippis, Evan/Radanovic, Goran/Parkes, David C./Liu, Yang (2020). How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. artificial intelligence 283. https://doi.org/10.1016/J.ARTINT.2020.103238.

Sloane, Stephanie/Baillargeon, Renée/Premack, David (2012). Do infants have a sense of fairness? Psychological Science 23 (2), 196–204. https://doi.org/10.1177/0956797611422072.

Srivastava, Megha/Heidari, Hoda/Krause, Andreas (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. Anchorage, AK, USA, Association for Computing Machinery, 2459–2468.

Suen, Hung-Yue/Chen, Mavis Yi-Ching/Lu, Shih-Hao (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? Computers in Human Behavior 98, 93–101.

van den Broek, Elmira/Sergeeva, Anastasia/and Huysman, Marleen (2019). Hiring Algorithms: An Ethnography of Fairness in Practice. ICIS 2019 Proceedings. 6.

van der Aalst, Wil M. P. (2017). Responsible data science: Using event data in a "People Friendly" Manner. In: Slimane Hammoudi/Leszek A. Maciaszek/Michele M. Missikoff et al. (Eds.). Enterprise Information Systems, Cham, 2017. Cham, Springer International Publishing, 3–28.

van Esch, Patrick/Black, J. Stewart/Ferolie, Joseph (2019). Marketing AI recruitment: The next phase in job application and selection. Computers in Human Behavior 90, 215–222. https://doi.org/10.1016/j.chb.2018.09.009.

Wang, Ruotong/Harper, F. Maxwell/Zhu, Haiyi (2020). Factors influencing perceived fairness in algorithmic decision-making. In: Regina Bernhaupt (Ed.). Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu HI USA. New York, NY, United States, Association for Computing Machinery, 1–14.

Wexler, J./Pushkarna, M./Bolukbasi, T./Wattenberg, M./Viégas, F./Wilson, J. (2020). The What-If Tool: Interactive Probing of Machine Learning Models. IEEE Transactions on Visualization and Computer Graphics 26 (1), 56–65. https://doi.org/10.1109/TVCG.2019.2934619.

Zeileis, Achim/Köll, Susanne/Graham, Nathaniel (2020). Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. Journal of Statistical Software 95 (1). https://doi.org/10.18637/jss.v095.i01.