AI and Ethics (2026) 6:29
https://doi.org/10.1007/s43681-025-00854-x

RESEARCH

Utility on the brain: an empirical investigation of fairness perceptions of algorithmic decisions under a utility-based ethical evaluation framework

Serhiy Kandul[1] · **Corinna Hertweck**[1] · **Ulrich Leicht-Deobald**[2]

Received: 3 September 2025 / Accepted: 23 October 2025

Abstract

This paper presents insights from a controlled experiment that examines how exposure to a utility-based fairness evaluation framework affects individual's perceptions of fairness and preferences for various fairness metrics in algorithmic allocation decisions. The framework's key features are (1) the explicit reference to the otherwise implicit utility assumptions for the affected groups and (2) the inclusion of deservedness arguments, a central tenet from moral philosophy. Conducted in the context of an employment agency that allocates job coaching slots, the study investigates two core questions: (1) How does the utility-based evaluation framework influence preferences for fairness metrics? (2) How does it affect the fairness perceptions of the resulting outcomes? The results show that people exposed to the utility-based evaluation framework choose different fairness metrics than people without such ethical guidance.

Keywords AI fairness · Fairness metrics · Utility-based framework · Fairness perceptions

1 Introduction

The increasing use of algorithms to make allocation decisions in various contexts, such as personnel selection, lending, and law enforcement, has raised important questions about fairness Rabonato and Berton [1], Shams et al. [2]. How should fairness be defined and measured in such decision-making processes?

Fairness is a multifaceted concept that is difficult to assess in practice, particularly when decision-making processes involve artificial intelligence (AI). Some argue that the involvement of AI for certain types of decisions renders the process inherently unfair or at least ethically problematic (*procedural fairness*, AI ethics; e.g., Newman et al. [3], Mittelstadt et al. [4], Vanderelst and Winfield [5]), while others emphasize the outcomes of human versus AI decisions and the potential for *algorithmic bias* (e.g., Cossette-Lefebvre and Maclure [6], Garcia et al.[7]). Defining and measuring AI fairness remains a substantial challenge because individuals hold diverse intuitions about what constitutes fairness, and because the inclusion of AI adds additional layers of complexity. To foster a more informed discussion on this issue, there is an urgent need for a conceptual framework that allows systematic evaluation of fairness in AI-driven decision-making. Although several such frameworks and tool-kits for evaluating the fairness of AI models have been proposed (e.g., Carey and Wu [8], Bellamy et al. [9], Bird et al. [10]), empirical evidence on how they shape people's perceptions of fairness–and, in particular, whether they provide sufficient guidance amid the myriad of proposed *fairness metrics*–remains limited. Our study takes a step toward addressing this gap by investigating how structured exposure to *fairness metrics*–through a component[1] of the conceptual evaluation framework presented in [11, 12]–affects individuals' preferences for fairness metrics and their perceptions of fairness in allocation outcomes.

✉ Serhiy Kandul
serhiy.kandul@ibme.uzh.ch

Corinna Hertweck
corinna.hertweck@uzh.ch

Ulrich Leicht-Deobald
ulrich.leicht-deobald@tcd.ie

[1] University of Zurich, Zurich, Switzerland

[2] Trinity College Dublin, Dublin, Ireland

[1] We exclude the distributive pattern component, which was revised in [37].

Published online: 01 December 2025

This evaluation framework, rooted in philosophical theories of justice, consists of four key components:

1. Utility of the decision subjects: Defines how to quantify the benefits/harms for decision subjects.
2. Relevant groups: Defines the social groups to be compared with respect to how much utility they receive.
3. Claim differentiator: Defines the characteristics that justify inequalities in utility distribution between individuals.
4. Pattern of justice: Defines what constitutes a just distribution.

In this manuscript, we aim to examine the extent to which people prefer different fairness metrics in an experimental setting and how these preferences are shaped by (1) utility considerations (2) of demographic groups, for which we focus on subgroups defined by (3) deservedness considerations under (4) an egalitarian pattern of justice.

Our primary research questions are as follows:

- How does exposure to the utility-based framework influence participants' preferences for fairness metrics?
- How does the utility-based framework affect participants' perceptions of fairness in allocation outcomes?

The experiment is framed around a scenario of an employment agency that allocates spots in a job coaching program to unemployed individuals, using predictions of the likelihood of finding a job based on participant attributes and race as a sensitive attribute to define relevant social groups. When considering the utilities of the job seekers, we also highlighted the *cost-effectiveness* of the job coaching program, which could possibly justify inequalities across the demographic groups.

Our paper offers an important theoretical contribution to the literature on algorithmic fairness as it theorizes and demonstrates that proper consideration of utilities and deservedness arguments shifts people's fairness preferences and perceptions. Hence, our manuscript demonstrates that an ethical guidance framework enables lay people to make more informed decisions regarding fairness metrics. As such, our framework allows stakeholders to arrive at more morally justified choices, which eventually should lead to a higher social acceptability of prediction-based decision-making systems.

2 Related work

A growing body of research examines the question of fairness in algorithmic decision-making, with many studies proposing different frameworks for evaluating fairness through so-called fairness metrics. One major area of focus is how individuals understand and evaluate these fairness metrics, as well as how they perceive fairness of algorithmic decisions more broadly.

Our study falls within the realm of 'algorithmic predictors' of fairness perceptions, as defined by the review work on algorithmic fairness by Starke et al. [13]. 'Algorithmic predictors' are features of algorithms that affect fairness perceptions of algorithmic decisions. This broad class of factors includes features such as impartiality of algorithms, the data they are trained on, and even the presence of algorithms per se in the decision-making process. Apart from controlling for the participants' age and gender, we do not explicitly include "human predictors" (Starke et al. [13], i.e., the factors on the participants' side that might influence fairness judgments (for a discussion on how participants' demographic features affect fairness perceptions, see, e.g., Grgić-Hlača et al. [14]). However, our paper contributes to a subfield of studies exploring people's preferences for *distributive fairness* of algorithms based on a particular set of *statistical measures* or fairness metrics for binary classifications.[2]

In the algorithmic fairness literature, it is well known that fairness criteria are often conflicting, which means that they cannot be fully fulfilled simultaneously in most cases in practice see Kleinberg et al. [15], Chouldechova [16], Barocas et al. [17]. Given that some fairness criteria seem more intuitively appealing to practitioners than others, the choice between these criteria is often difficult in practice, because each fairness criterion represents different values and a different view of what it means to be fair [18]. As such, some deliberation and guidance are needed in practice to select the most appropriate fairness metric. However, for practitioners, obtaining this guidance can be challenging. Fairness toolkits, such as AIF360 [9] or Fairlearn [10], can help practitioners evaluate the fairness of their model by providing an implementation of various fairness criteria. However, interviews and surveys with practitioners revealed that practitioners are often overwhelmed by the sheer number of fairness metrics presented to them and lack guidance on how to choose among them [19–21]. Specifically, so far there are few accessible explanations in those toolkits as to when and how to use the implemented metrics [21]. This situation makes it difficult for practitioners to make an informed

[2] For research on perception of fairness metrics for non-binary classifications in the context of scholarship allocations, see e.g., Alkhathlan et al. [38].

decision. Our study examines whether and how a guiding framework that reframes fairness criteria through the lens of utility and deservedness considerations affects people's fairness choices and perceptions.

In doing so, we draw upon the literature that studies people's preferences regarding conflicting fairness metrics. This body of work suggests that people's preferences over fairness metrics are highly context-specific. Srivastava et al. [22] compared a few prominent mathematical definitions of fairness metrics with human decision makers' preferences in the context of criminal risk and cancer risk predictions. The authors demonstrate that in both contexts, people's choices are predominantly consistent with "simpler metrics," such as statistical parity. In the context of bail, Harrison et al. [23] observed that a slim majority of participants preferred to equalize false positive rates (i.e., the frequency of mistakenly defined bails) over equalizing accuracy and statistical parity among white and African-American groups of defendants. However, a substantial minority preferred accuracy and statistical parity. In the context of pre-screening job candidates, Sengewald et al. [24] found that automated decision-making systems constrained by equality of opportunity and equalized odds (i.e., the share of qualified candidates invited for the interview and the share of unqualified candidates invited for the interview between the two groups) are perceived to be more fair than those optimized for statistical parity. The mixed results suggest that people's attitudes towards fairness metrics are context-dependent and sensitive to the methods used to elicit them (e.g., the differences emerge between explicit presentation of the trade-offs between metrics or inferred metrics from the model output).

In sum, existing empirical research on participant's preferences documents people's preferences over fairness metrics "in the wild", i.e., with minimal provision of any moral or ethical guidance to the participants. Our contribution is to measure whether people's preferences change when being prompted to evaluate the consequences of (we refer to these consequences as "utilities") and the deservedness of the individuals affected by the algorithmic decisions. Our framework encourages participants to (1) explicitly consider how the decision-making system impacts the relevant social groups (i.e., their *expected utilities*) and to 2) explicitly consider *deservedness* arguments which define the relevant subgroups for utility comparisons (and justify inequalities between the groups when all group members are considered).[3]

On a conceptual level, our approach is related to studies that map utility-based moral evaluation to a specific class of statistical notions of fairness. Heidari et al. [25], for example, introduced an ethical utility-based framework for fairness that maps different moral viewpoints, such as Rawlsian and luck egalitarian accounts of equality of opportunity, with different fairness criteria. Loi et al. [26] discuss utilities as an ethical principle that guides the choices of fairness metrics.

The concept of *deservedness* in fairness is also deeply rooted in moral philosophy. Philosophers like Feldman [27] and Miller [28] discussed desert as one of the key principles of justice, clarifying the role of individual responsibility and contribution as a basis upon which individuals are deemed deserving of particular outcomes. Empirical studies in economics and psychology confirm the profound role of accountability in perceptions of desert, alongside other related principles such as merit, effort, or need, in defining people's fairness perceptions and preferences for redistribution Konow [29], Tinghög et al. [30], Cappelen et al. [31]. In these studies, merit is typically operationalized by allowing participants to perform a task and linking participants' payment to their performance. Need is operationalized as a loss of income due to an unfavorable random shock ("brute luck"), which reduces the reward of a participant and is beyond the participant's control (see, for example, the solidarity game in Selten and Ockenfels [32]). In the algorithmic fairness literature, need is sometimes simply portrayed as the low level of income of the recipients (see, for example, Lee et al. [33]). Finally, accountability is manipulated as "option luck". In settings with option luck, the initial endowment of participants is based on their previous choices, e.g., the choice to invest in a risky option (and later lose the lottery) as in Mollerstrom et al. [34]. After the initial income distribution and the source of inequalities are disclosed to the participants, they can redistribute payoffs, either as affected parties or as impartial spectators. The degree of redistribution then reflects the weight people assign to the sources of inequality.

Although empirical evidence suggests that people tend to differentiate between brute luck and option luck in generic allocation scenarios – for example, sharing less money with recipients who lost income due to their own choices – demarcating the link between luck, choice, and deservedness in more complex environments is challenging [35]. Binns [35] discusses cases in which individuals make seemingly disadvantageous choices, such as self-selecting into low-paid jobs or moving into high-crime neighborhoods, which may nonetheless yield societal benefits (e.g., reducing healthcare costs for dependents or contributing to community security). Binns argues that some form of compensation for these individual choices is still justified and calls for a broader consideration of the societal implications of people's decisions when assessing deservedness.

[3] Although scenarios in existing studies typically contain information on benefits or harms for the affected individuals, to the best of our knowledge, there are no explicit references or requests to compute and compare utilities or deservedness across the relevant subgroups.

In our study, we add to the discussion of perceptions of deservedness in the context of allocating job coaching. In particular, we ask participants to consider *cost-effectiveness* as a justification for inequalities and for a separate consideration of the utilities of the subgroups based on the true labels. We focus on situations where true labels $Y = 1$ imply the better use of resources (and at the same time potentially result in higher utility for affected individuals). In our scenario, the allocation of job candidates who can find a job after coaching ($Y = 1$) guarantees that the program's costs are not "wasted" on those who cannot find a job ($Y = 0$). Relying on evidence from economic experiments about efficiency-seeking as a potential motive behind redistribution Konow [36], Charness and Rabin [29], we believe that cost-effectiveness can be an essential factor in people's attitudes towards fairness metrics. Lee et al. [33] provide empirical evidence that people care about efficiency (delivery costs to food recipients) for algorithm-based decisions in the context of donor-recipient matching algorithms. Therefore, the literature suggests that cost-effectiveness is an important factor for people's fairness evaluations.

Overall, our key contribution to the literature on algorithmic fairness is the empirical investigation of participants' preferences over conflicting fairness metrics and their fairness perceptions of the allocation decisions in the context of (un)employment when being prompted to consider the utilities of the job seekers and deservedness arguments based on cost-effectiveness.

3 Experimental design

3.1 Scenario description

The experiment involves participants acting as decision-makers in a simulated employment agency scenario. The agency must allocate limited spots in a job coaching program to unemployed individuals based on model predictions of their ability to find a job after completing the program.

Participants are presented with a two-year time horizon. In the first year, individuals either join the coaching program or do not. Success or failure to find a job is recorded in the second year. Job seekers can find a job after the coaching program (true $Y = 1$) or not (true $Y = 0$). For simplicity, we assumed that no job seeker can find a job without a job coaching program. The employment agency covers participation costs in the job coaching program (USD 10K). It pays a fixed amount of unemployment benefits (USD 75K) in Year 1 and Year 2 for everyone unemployed in the respective year. Finding a job implies receiving a salary of USD 150K (twice as high as the unemployment benefits), i.e., an increase in the well-being of a job seeker. The employment agency tries to predict Y (we denote the prediction as $\hat{Y}$) and allocates everyone with $\hat{Y} = 1$ to the job coaching program. The agency suffers a financial loss for misallocations (sending people with true $Y = 0$ to the program, or not sending people with true $Y = 1$ to the program).

We use race as the sensitive attribute in this experiment, from which we construct two demographic groups, Black and white job seekers. This oversimplified assumption allows us to test fairness perceptions in the simplest case of just two demographic groups. Although some scholars argued for more neutral labels of the groups (like orange vs. blue group) to mitigate potential in-group bias in fairness judgments (Sengewald et al. [24], using race as a sensitive attribute is still common in empirical studies on fairness perceptions of algorithmic decisions, see e.g. Srivastava et al. [22], Harrison et al. [23]. Although this design choice might have inflicted some bias in people's fairness judgments, race definitely served as a "relevant" social comparison.

The key fairness metrics evaluated in the experiment are:

- *Accuracy Comparison (AC)*: The proportion of individuals with correct predictions (true positives and true negatives) among the total population across the two demographic groups.
- *Statistical Parity (SP)*: The proportion of individuals allocated to the program among the total population across the two demographic groups.
- *Equality of Opportunity (EO)*: The proportion of individuals who are correctly allocated to the program ($\hat{Y} = 1$ and $Y = 1$) among those with a positive true outcome ($Y = 1$) across the two demographic groups.
- *True Positives Comparison (TPC)*[4]: The proportion of individuals who are correctly allocated to the program ($\hat{Y} = 1$ and $Y = 1$) among the total population across the two demographic groups.

The literature on algorithmic fairness proposes a wide range of fairness metrics, each addressing different notions of fairness. Statistical parity (SP) and Equality of Opportunity (EO) represent some of the most prominent classes of fairness metrics used in the field. These metrics capture different perspectives on fairness and reflect important trade-offs in real-world decision-making contexts.

In our setting, the algorithmic decisions always disadvantaged the historically and systematically discriminated Black candidates. Importantly, fairness metrics produced

[4] Note that this metric is not established in the literature but is rather derived for this scenario using our utility-based framework. Therefore, the name of this metric is also not an established term but rather one we use to refer to the metric in this paper only. The derivation process for this metric is explained further down in the description of the treatment condition in DV2.

various degrees of disparities among the Black and white job seekers. The Total Accuracy of the model was 77% and 72% for white and Black job seekers, respectively. An evaluation using Statistical Parity showed that 88% of white job seekers and 67% of Black job seekers were given a positive decision (allocated to a job coaching program). Using Equality of Opportunity to evaluate fairness, we saw that 78% of white job seekers and 69% of Black job seekers who would find a job with the help of the program were allocated to the job coaching program. Finally, a comparison of the share of the true positives compared to the total population in both groups (we called this the "True Positives Comparison") gives us a value of 0.69 for white job seekers and only 0.46 for Black job seekers. Thus, the Accuracy Comparison and Equality of Opportunity produced smaller disparities among the demographic groups than the other two fairness metrics.

We then asked participants to choose the fairness criterion they deem most appropriate in the context and elicit their fairness perceptions of the job allocation decisions before and after the choice of a particular fairness metric.

3.2 Experimental procedure

We pre-registered the study.[5] The experiment was conducted in March 2024 at Laboratory for experimental and behavioral economics of the University of Zurich. We thoroughly followed our university's ethical guidelines for IRB approval. First, we obtained informed consent from the participants after they read the experimental instructions. Second, participants were instructed that their participation was voluntary and that they could withdraw from the study at any time. Based on our university guidelines, no further IRB approval was required, as our study did not impose any risks for participants or involve any forms of deception.

3.2.1 DV1: ex-ante fairness

Upon arrival at the lab, participants were randomly assigned to either a baseline or treatment condition. The experimental flow is presented in Fig. 1.

In both conditions, after reading the scenario, participants were first shown the information about the total number of unemployed individuals in each of the two demographic groups and how they were divided according to their true and predicted ability to find a job after the coaching program. This information was presented in the form of 2 by 2 distribution tables (see Fig. 2). Participants were then asked to evaluate the fairness of the job allocation decisions without any further guidance. We refer to this as **DV1**.

DV1: Ex-ante fairness

Participants rated the fairness of the allocation based on the information presented in the distribution tables, using a scale of 1 to 7.

We used subscales (1 to 7) to evaluate general fairness, fairness for Black candidates, and fairness for white candidates.

Figure 2 shows two tables illustrating how accurately the model predicted outcomes for white candidates (panel a) and Black candidates (panel b). Each table compares the actual number of people with ($Y = 0$ or $Y = 1$) shown in the columns with the model's predictions ($Y = 0$ or $Y = 1$) shown in the rows.

This layout allows readers to see both the correct and incorrect predictions at a glance. Participants also knew that everyone predicted as $Y = 1$ was selected for the program, and that only those whose true outcome was $Y = 1$ actually found a job and received a salary.

With this information, participants could consider the consequences of each possible situation:

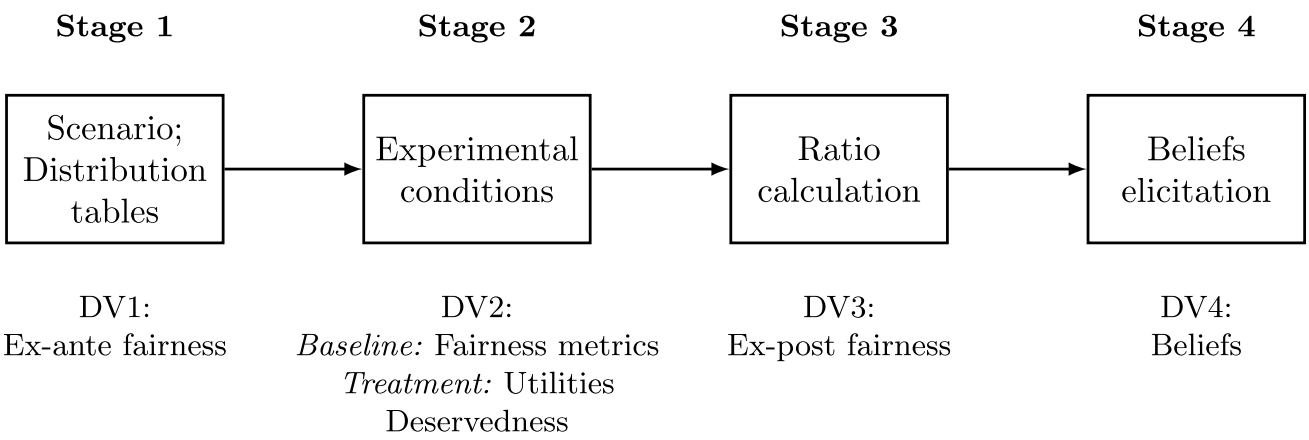

Fig. 1 Timeline of the experiment

[5] https://aspredicted.org/2kv8-n7vq.pdf.

Fig. 2 Distribution of accurate and inaccurate model predictions for white and Black participants

Model output: White participants

	True Y=0	True Y=1	Total
Predicted Y=0	987 (true negative)	437 (false negative)	1424
Predicted Y=1	2325 (false positive)	8202 (true positive)	10 527
Total	3312	8639	11 951

(a) White job seekers

Model output: Black participants

	True Y=0	True Y=1	Total
Predicted Y=0	249 (true negative)	75 (false negative)	324
Predicted Y=1	200 (false positive)	447 (true positive)	647
Total	449	522	971

(b) Black job seekers

1. Did not join the program and would not have found a job (predicted $Y = 0$, and true $Y = 0$, "true negative" prediction.
2. Did not join the program but could have found a job (predicted $Y = 0$, and true $Y = 1$, "false negative" error.
3. Joined the program but did not find a job (predicted $Y = 1$, and true $Y = 0$, "false positive" error.
4. Joined the program and found a job (predicted $Y = 1$, and true $Y = 1$, "true positive" prediction.

These distribution tables therefore provide all the information about the frequency of correct and wrong model predictions and enable judgments about the model's fairness. Although this was not said explicitly, we assumed that the group of Black candidates represented a marginalized demographic group because of the historic and systematic discrimination they experienced in the US, where our fictional example was set.

3.2.2 Control questions

Before proceeding to the next stage of the experiment, participants answered a few control questions targeting their understanding of types of errors, false positives and false negatives, through computational examples. The major goal was to establish a basic understanding of the distribution tables and the scenario (especially the cost of errors of each type).

3.2.3 DV2: fairness metric preference

Subsequently, participants were presented with our set of four fairness metrics as possible 'criteria' to evaluate the fairness of the job allocation decisions. The fairness metrics were therefore labelled neutrally, e.g., we never referred to "Equality of Opportunity' not to induce any sentiment towards this metric being more equal or more fair than

others. We just referred to comparisons of the respective ratios from the distribution tables.

Participants were asked to choose the one fairness metric out of the four offered ones that they deem to be *the most appropriate* in the given context. We refer to this part of the experiment as **DV2**.

DV2: Fairness metric preference
Participants selected the most appropriate fairness metric from a list of options.

The options were presented to them as follows:

- *Accuracy Comparison (labelled as "Criterion 1")*:"The comparison of the total accuracy for Black and white job seekers, i.e., the comparison of the ratio *(true positive + true negative)/total number of job seekers* across the two groups;
- *Statistical Parity (labelled as "Criterion 2")*: "The comparison of the total share of persons allocated to a program for Black and white job seekers, i.e., the comparison of the ratio *predicted* $Y = 1$*/ total number of job seekers* across the two groups;
- *Equality of Opportunity (labelled as "Criterion 3")*: "The comparison of the share of persons with true Y=1 who are assigned to the program out of the group with true Y=1 for Black and white job seekers, i.e., the comparison of the ratio *true positive/total number of people with* $Y = 1$ across the two groups;
- *True Positives Comparison (labelled as "Criterion 4")*: The comparison of the share of persons who get allocated to a program and eventually find the job for Black and white job seekers, i.e., the comparison of the ratio *true positive/total number of job seekers* across the two groups.[6]

Utility matrix: Option 1

	True Y=0	True Y=1
Predicted Y=0	U=0	U=0
Predicted Y=1	U=1	U=1

(a) Utility Matrix 1

Utility matrix: Option 2

	True Y=0	True Y=1
Predicted Y=0	U=0	U=0
Predicted Y=1	U=0	U=1

(b) Utility Matrix 2

Utility matrix: Option 3

	True Y=0	True Y=1
Predicted Y=0	U=1	U=0
Predicted Y=1	U=0	U=1

(c) Utility Matrix 3

Fig. 3 The set of utility matrices presented in the treatment condition

In the *baseline condition*, these four fairness metrics were described as above. Beyond that, participants were not given any guidance.

Contrary to the baseline condition, participants received further guidance in the *treatment condition*: To elicit preferences over fairness metrics in the treatment condition, participants were first exposed to utility matrices which summarize the relative gain in well-being for an individual in each possible situation (Fig. 3) and asked to choose the one they thought fit best the scenario. They were encouraged to think about the situation and select the matrix that, in their view, best captured how benefits in each possible outcome were distributed. Figure 3 presents the utility matrices used in the study. To keep the task simple, we only presented matrices with utility values of 0 (= no benefit in the given situation) and 1 (maximum benefit in the given situation).[7] Similar to distribution tables depicting correct and incorrect predictions of the model, each utility matrix specified a utility value for every possible combination of $\hat{Y}$ (whether the model predicts that the job seeker will find a job, which determines whether they are assigned to the program) and Y (whether the job seeker actually finds a job after the coaching program), i.e. four situations in total.

The different utility matrices represented alternative interpretations of the benefits and harms of the job coaching scenario. Utility matrix 1, for example, assumes an equal utility of $U = 1$ for everyone participating in the job coaching program. This implies that all participants are considered to benefit from the program regardless of whether

[6] The options were presented into two possible orders (random assignment, between-subject) which differ in the first and the last option presented (with Criterion 1 and Criterion 4 flipped), i.e. (Accuracy Comparison, Statistical Parity, Equality of Opportunity, and True Positives Comparison) or (True Positive Comparison, Statistical Parity, Equality of Opportunity, and Accuracy Comparison).

[7] Participants who believed that there was "some" but smaller benefit (i.e., a utility gain between 0 and 1) had therefore to consider whether it was closer to 0 or 1.

they ultimately find a job. One possible reasoning is that participation in the program still provides valuable skills, experience, or confidence gains. Utility matrix 2, on the other hand, assumes that individuals who do not find a job after completing the program do not experience a meaningful increase in utility (i.e., their well-being remains similar to that of those who only receive unemployment benefits)[8] In this matrix, therefore, only those who participate in the program and subsequently find a job have a utility of $U = 1$. We hypothesize that this utility matrix best reflects the scenario and would therefore be the most popular choice among participants. Within the framework that we built on, which outlines the procedure for defining a utility matrix in [12], utility matrix 3 presents a less plausible interpretation.[9] This matrix was nevertheless included to ensure comparability with the baseline condition, since it is the only matrix that corresponds to the accuracy-based fairness metric. For this reason, its inclusion is primarily technical, and we expected it to be chosen least often.

Utility matrices, therefore, do not add any new relevant information beyond the consequences for job seekers in each of the four possible situations described in the scenario. Instead, they allow for different interpretations of these consequences, i.e. their mapping to the change of the well-being of the affected individuals. We did not provide further guidance on which utility matrix to choose, apart from encouraging participants to think carefully about the scenario.

After participants selected their preferred utility matrix, we introduced the concept of *deservedness*. Participants were asked whether they believed that a person's deservedness to receive job coaching should depend on whether that person ultimately finds a job after the coaching. Cost-effectiveness was presented as a possible rationale for why some individuals might be viewed as more deserving of participation than others. Specifically, participants were presented with the following instructions: "To evaluate fairness, please consider the average utility of black and white job candidates. When doing so, one might say that all long-term unemployed people in both groups deserve the same (average) utility from the program, regardless of whether they find a job at the end (regardless of their true *Y*). Others might argue that not everyone *deserves* the same utility from the program, e.g., because spending the costs of the program on people who do not find a job after the coaching is not justified. Based on these considerations, we next present a selection of possible criteria for fairness evaluation."

To keep the concept of deservedness straightforward, participants were presented with two options:

1. Compare the utility of **all** job seekers across both groups (including those with $Y = 0$ and $Y = 1$).
2. Compare the utility of only those who **actually found a job** ($Y = 1$) across both groups.

Choosing the first option (considering both $Y = 0$ and $Y = 1$) indicates the belief that all individuals are equally deserving of participation in the job coaching program. Choosing the second option (only true $Y = 1$) reflects the belief that participation should depend on outcomes–such as job attainment–possibly for reasons of cost-effectiveness.

However, these two options were not presented to them disconnected from the utility matrix as above. Rather, participants were shown two options, formulated based on the utility matrix they had chosen: If a participant had chosen utility matrix 1, they were asked to choose between:

- "Criterion 1 (same average utility for all): The comparison of the total share of persons allocated to a program for black and white job seekers, i.e., the comparison of the ratio *predicted* $Y = 1$/ *total number of job seekers* across the two groups;" and
- "Criterion 2 (same average utility for $Y = 1$): The comparison of the share of persons with true $Y = 1$ who are assigned to the program out of the group with true $Y = 1$ for black and white job seekers, i.e., the comparison of the ratio *true positive/total number of people with* $Y = 1$ across the two groups."

If a participant had chosen utility matrix 2, they were asked to choose between:

- "Criterion 1 (same average utility for all): The comparison of the total share of persons allocated to a program who actually find a job for black and white job seekers, i.e., the comparison of the ratio *true positives/total number of job seekers* across the two groups;" and
- Criterion 2 (same average utility for $Y = 1$): Same as criterion 2 above.

Finally, if a participant had chosen utility matrix 3, they were asked to choose between:

- "Criterion 1 (same average utility for all): The comparison of the total accuracy for black and white job seekers, i.e., the comparison of the ratio *(true positive+true*

[8] Because we use binary utilities, this means the program's benefit is considered negligible if participants do not find a job; the main benefit comes from successful employment.

[9] A possible justification is that job seekers with $Y = 0$ with pro-social preferences might appreciate not "taking the place of" of someone who could have benefited more. However, this assumes that people know about their true *Y*, which they are, of course, actually unaware of.

negative)/ total number of job seekers across the two groups;" and
- Criterion 2 (same average utility for $Y = 1$): Same as criterion 2 above.

Importantly, the combination of the choice of the utility matrix and the choice of the deservedness option in the treatment condition lead to the same set of four fairness metrics as in the baseline.

- *Accuracy Comparison*: Utility matrix 3 and deservedness option 1;
- *Statistical Parity*: Utility matrix 1 and deservedness option 1;
- *Equality of Opportunity*: Any of the utility matrices and deservedness option 2;
- *True Positives Comparison*: Utility matrix 2 and deservedness option 1.

Although the elicitation of preferred fairness metrics differed between experimental conditions, participants in both settings were aware of the consequences of correct and incorrect model classifications. To ensure this understanding, they were asked control questions immediately before DV2. Thus, participants in the baseline condition could also consider the associated utilities implicitly.

3.2.4 Ratio calculation

After the participants' preferred fairness metric were identified, we asked them in both experimental conditions to compute the respective ratios (based on the confusion matrices). For this task, we showed participants the confusion matrices again (and for the treatment condition, also the chosen utility matrix) and asked them to compute the respective ratios for the fairness metric of their choice.[10]

3.2.5 DV3: ex-post fairness

Finally, once participants had computed the respective ratios for the Black and white job seekers and therefore learned the underlying degree of disparity, we asked them again how fair they found the allocation decision in **DV3**. We used subscales to evaluate general fairness, fairness for Black candidates, and fairness for white candidates.

DV3: Ex-post fairness

Participants rated fairness again after computing ratios based on their chosen metric on a scale from 1 to 7.

[10] Participants could not proceed to the next screen until they arrived at correct numbers.

3.2.6 DV4: Participant's beliefs

Finally, we elicited participants' incentivized beliefs about the fairness metrics most frequently chosen by others. In the baseline, participants were simply asked which of the four fairness metrics they believed was the most frequently chosen by participants in their experimental sessions. In the treatment, participants were asked which of the utility matrices and deservedness arguments they believed were the most frequently chosen by participants in their experimental session.

4 Results

In total 109 participants took part in the experiment conducted at the Laboratory of experimental and behavioral economics at the University of Zurich in March 2024. The mean age of the participants was 23; around 40% identified as women. All participants had a background in STEM (science, technology, engineering and mathematics). A session lasted approximately 1 h and 15 min; the average payment was 30.6 CHF.

4.1 DV1: ex-ante fairness perceptions

We begin by comparing the ex-ante fairness perceptions, i.e. the judgment of fairness of the allocation of Black and white unemployed people to the job coaching program based on the information presented in Stage 1 of the experiment.

The pairwise comparisons (t.test with Bonferroni correction) indicate significant differences across the fairness subscales:

fair_general vs fair_black (Stage 1): $t(216.0) = -2.61$, $M_{fair\_general}$=3.20, $SD_{fair\_general}$=1.36, $n_{fair\_general}$=109; M_{fair_black}=3.54, SD_{fair_black}=1.59, n_{fair_black}=109; p=0.010; *Cohen's d* = –0.35;

fair_general vs fair_white (Stage 1): $t(216.0) = -7.08$, $M_{fair\_general}$=3.02, $D_{fair\_general}$=1.36, $n_{fair\_general}$=109; M_{fair_whitel}=4.48, D_{fair_whitel}=1.66, n_{fair_whitel}=109; p< 0.001; *Cohen's d* =–0.96;

fair_black vs fair_white (Stage 1): $t(216.0) = -4.24$, M_{fair_black}=1.54,SD_{fair_black}=1.59,n_{fair_black}=109;M_{fair_white}=4.48, SD_{fair_white}=1.66, n_{fair_white}=109; p< 0.001; *Cohen's d* =–0.96.

For ex-ante fairness evaluations in DV1, participants indicated that they found the allocation decisions fairer for white candidates. This is expected, as by design, the algorithmic decisions were more advantageous for white participants.

We observed no differences in ex-ante fairness evaluations between the baseline and treatment condition (pairwise

t.test comparisons with Bonferroni correction, $p > 0.10$) for all the three fairness scales:

fair_general (Stage 1): $t(107.0) = -1.41$, $M_{base} = 2.83$, $SD_{base} = 1.13$, $n_{base} = 54$; $M_{treat} = 3.20$, $SD_{treat} = 1.54$, $n_{treat} = 55$; $p = 0.161$; $Cohen's\ d = -0.27$;

fair_black (Stage 1): $t(107.0) = -1.61$, $M_{base} = 3.30$, $SD_{base} = 1.57$, $n_{base} = 54$; $M_{treat} = 3.78$, $SD_{treat} = 1.58$, $n_{treat} = 55$; $p = 0.111$; $Cohen's\ d = -0.31$;

fair_white (Stage 1): $t(107.0) = -1.48$, $M_{base} = 4.24$, $SD_{base} = 1.84$, $n_{base} = 54$; $M_{treat} = 4.71$, $M_{treat} = 1.45$, $n_{treat} = 55$; $p = 0.143$; $Cohen's\ d = -0.28$.

This is as expected since no experimental manipulation has been introduced at this stage of the experiment.

4.2 DV2: Fairness metric preference

We now turn to the choices of fairness metrics after the experimental manipulation has been introduced.

4.2.1 Treatment condition

First, let us start with the choices of the utility matrix and the deservedness argument, which implicitly defines the choice of fairness criterion in the treatment. 25 out of 55 participants, i.e. around 45% of participants, chose utility matrix 2 (see Fig. 3b), which assigned a utility value of $U = 1$ for job seekers with true $Y = 1$ and predicted $\hat{Y} = 1$ and $U = 0$ to all other cases. This implies that participants believed that the job seekers only derive a positive utility if they are able to find a job after the job coaching program and were in fact allocated to participate in the program — in line with our hypothesis that this is the most plausible utility matrix. Utility matrix 1 and 3 were each chosen by 15 out of 55 participants, i.e., by around 27%. Utility matrix 1 (see Fig. 3a) implied that the allocation to the job coaching program also benefited participants who were not able to find the job afterwards, and utility matrix 3 (see Fig. 3c) implied a positive utility for a correct prediction only. The frequency at which utility matrix 3 was chosen — despite being lower than for utility matrix 2 — was higher than we initially expected. However, given that participants were, on purpose, given minimal instructions so as not to influence them in their ethical evaluation of the scenario, it is less surprising that people interpreted utility matrices differently from us.[11] Although more participants chose the utility matrix 2, which is consistent with our hypothesis, the distribution of choices of the utility matrices is not statistically different from the uniform distribution ($p = 0.20$, exact multinomial test; Cramer's $V = 0.18$), consistent with either "random decisions" (indifference among matrices) or an equal share of participants who strongly preferred each option.

With respect to deservedness, slightly more than a half of participants, 31 out of 55 participants or about 56%, have agreed that the fairness evaluation should take Y into account: 7 of 15 of those who had chosen utility matrix 1, 15 out of 25 of those who had chosen utility matrix 2, and 9 out of 15 of those who had chosen utility matrix 3. Fisher exact test indicates that the share of those who chose the deservedness option 2, i.e., to include Y in the evaluation, is not statistically different from 50%, $p = 0.77$. The association between utility matrix and deservedness choice was small, Cramer's $V = 0.12$, suggesting a weak link between the assigned matrix and participants' fairness evaluation. This pattern is consistent with a "random draw" choice, reflecting indifference among options, or an equal distribution of strong preferences.

4.2.2 Baseline condition

Let us now compare the chosen metrics in the treatment condition (i.e., through the choice of a utility matrix and a deservedness argument) with those chosen in the baseline condition. We observed significant differences in distributions of the chosen fairness metrics across experimental conditions (see Fig. 4).

As Fig. 4 shows, the True Positives Comparison (TPC, fairness criterion 4) is the least preferred fairness metric in the baseline; the distribution of the chosen metric is significantly different from the uniform distribution ($p < 0.01$, exact multinomial test, n=54). This result suggests that, without any additional guidance, this criterion is

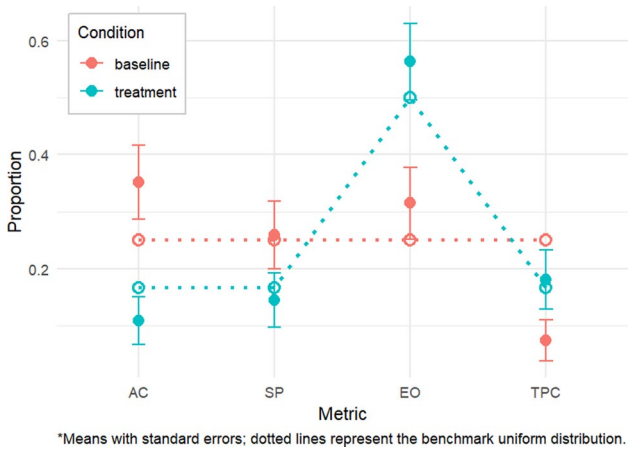

Fig. 4 Proportions of metrics chosen across conditions

[11] We elicited participants' overall open-ended explanations for their choices in the experiment. A few participants who chose utility matrix 3 stated they that correct classification increased well-being for job seekers with $Y = 0$ without providing a specific source of this well-being improvement.

not the most intuitive choice people would make.[12] As we saw in the treatment condition, people become more likely to choose this metric, though, when instructed to look at it through the lens of utilities. Specifically, each of the fairness metrics, including the True Positives Comparison, was equally preferred in the treatment condition ($p = 0.64$, exact multinomial test of the observed distribution to the expected distribution of (1/6, 1/6, 3/6, and 1/6) for the four fairness metrics, N=55, respectively, assuming that each utility matrix and each of the deservedness argument are equally likely to be chosen). Two-sample test of proportions indeed confirms a significant difference in the frequency of the choice of the True Positives Comparison (fairness criterion 4) as the fairness metric between the treatment and the baseline (z-score of 3.28).

We conclude that the introduction of utility matrices and deservedness arguments led to an increasing preference for the fairness metric that best aligns with a utility-based view of fairness metrics (i.e., the fourth fairness criterion, True Positives Comparison).

4.3 DV3: ex-post fairness & the role of disparities among the groups

We conclude by analyzing the ex-post fairness judgments, i.e., people's perception of fairness of the allocations to the job coaching program after they have chosen the fairness metric (the lens) and computed the resulting disparities across demographic groups.

Comparisons of ex-post fairness perceptions did not reveal any meaningful difference across the conditions. Mixed ANOVA reveals no significant difference for the experimental conditions:$F(1.00, 107) = 1.03$, $p = 0.313$, $\eta_g^2 = 0.006$, but significant differences for fairness subscales: $F(1.65, 177) = 140.58$, $p =< 0.001$, $\eta_g^2 = 0.318$ (and no significant interactions effects: $F(1.65, 177) = 0.11$, $p = 0.856$, $\eta_g^2 = 0.000$).

Similar to the ex-ante fairness perceptions, participants evaluated the job allocation decision as fairer for the white participants.

fair_general vs fair_black (Stage 3): $t(216) = -0.98$, $M_{fair\_general}$=3.2, $SD_{fair\_general}$=1.44, $n_{fair\_general}$=109; M_{fair_black}=3.39, SD_{fair_black}=1.46, n_{fair_black}=109; p=0.982; *Cohen's d*=–0.13.

fair_general vs fair_white (Stage 3): $t(216) = -10.29$, $M_{fair\_general} = 3.39$,$SD_{fair\_general}$=1.46, $n_{fair\_general}$=109; M_{fair_white}=5.27, SD_{fair_white}=1.21, n_{fair_white}=109; p=<0.001; *Cohen's d*=–1.39.

fair_black vs fair_white (Stage 3): $t(216) = -11.45$, M_{fair_black}=3.2, SD_{fair_black}=1.44, n_{fair_black}=109$n_{fair_black} = 109$; M_{fair_white}=5.27, SD_{fair_white}=1.21, n_{fair_white}=109; p=<0.001; *Cohen's d*=–1.55.

Interestingly, ex-post fairness evaluations are influenced by the choice of the fairness metric (general fairness,[13] conditions pooled). In particular, for the chosen metric of total model accuracy across the Black and white job seekers, the perceived fairness is rated as higher than under Statistical Parity or under the True Positives Comparison:

AC vs SP: $t(41) = 2.96$, $M_{\text{AC}} = 4.2$, $SD_{\text{AC}} = 1.29$, $n_{\text{AC}} = 25$,$M_{\text{SP}} = 2.95$,$SD_{\text{SP}} = 1.56$,$n_{\text{SP}} = 22$,$p = 0.031$, $Cohen'sd = 0.88$;

AC vs TPC: $t(36.7) = 5.46$, $M_{\text{AC}} = 4.2$, $SD_{\text{AC}} = 1.29$, $n_{\text{AC}} = 25$, $M_{\text{TPC}} = 2.5$, $SD_{\text{TPC}} = 0.65$, $n_{\text{TPC}} = 14$, $p < 0.001$, $Cohen'sd = 1.53$;

AC vs EO: $t(54.7) = 2.28$, $M_{\text{AC}} = 4.2$, $SD_{\text{AC}} = 1.29$, $n_{\text{AC}} = 25$,$M_{\text{EO}} = 3.44$,$SD_{\text{EO}} = 1.47$,$n_{\text{EO}} = 48$,$p = 0.159$, $Cohen'sd = 0.54$.

The other pairwise comparisons do not reveal significant differences:

SP vs EO: $t(38.8) = -1.23$, $M_{\text{SP}} = 2.95$, $SD_{\text{SP}} = 1.56$, $n_{\text{SP}} = 22$, $M_{\text{EO}} = 3.44$, $SD_{\text{EO}} = 1.47$, $n_{\text{EO}} = 48$, $p = 1$, $Cohen'sd = -0.32$

SP vs TPC: $t(30.4) = 1.21$, $M_{\text{SP}} = 2.95$, $SD_{\text{SP}} = 1.56$, $n_{\text{SP}} = 22$,$M_{\text{TPC}} = 2.5$,$SD_{\text{TPC}} = 0.65$,$n_{\text{TPC}} = 14$,$p = 1$, $Cohen'sd = 0.35$

EO vs TPC: $t(50) = 3.42$, $M_{\text{EO}} = 3.44$, $SD_{\text{EO}} = 1.47$, $n_{\text{EO}} = 48$, $M_{\text{TPC}} = 2.5$, $SD_{\text{TPC}} = 0.65$, $n_{\text{TPC}} = 14$, $p = 0.008$, $Cohen'sd = 0.7$

Remember that by design, our fairness metrics are associated with different levels of resulting disparities between the Black and the white candidates.

To analyze whether the degree of disparities affects the fairness evaluations, we run a linear regression:

$$\begin{aligned} Y_i =& b_0 + b_1 * Disparity_i \\ &+ b_2 * FairMetric_i + b_3 * Ctrls + e_i \end{aligned} \quad (1)$$

where Y_i is the ex-post fairness score from 1 to 5 assigned by a participant *i*, $Disparity_i$ is the absolute difference in the respective ratios for white and Black job seekers under the respective fairness metric, and $FairMetric_i$ is the fairness metric chosen by a participant *i*, *Ctrls* include age and gender, and finally e_i is the error term.

Table 1 presents results of the regressions run for general fairness, fairness for white, and fairness for Black job seekers (experimental conditions pooled).

[12] Interestingly, in contrast to [22], statistical parity was not the most frequent choice in any of our experimental conditions.

[13] As opposed to fairness for white job seekers and fairness for Black job seekers.

Table 1 Regression Results

	Fairness general	Fairness white	Fairness black
(Intercept)	7.253^{***} (1.125)	5.430^{***} (0.999)	7.417^{***} (1.085)
Disparity	-6.371^{**} (2.102)	-0.926 (1.865)	-7.550^{***} (2.026)
FairMetric	-0.172 (0.148)	-0.052 (0.131)	-0.102 (0.143)
Age	-0.094^{*} (0.046)	0.041 (0.040)	-0.109^{*} (0.044)
Gender	-0.353 (0.262)	-0.609^{*} (0.233)	-0.379 (0.253)
R^2	0.186	0.074	0.221
Adj. R^2	0.155	0.038	0.191
Num. obs	109	109	109

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

The results of the regression analysis reveal that disparity is a substantial and statistically significant negative coefficient for the general fairness perception and the perception of fairness for Black job seekers. The choice of fairness criterion is, however, an insignificant coefficient for the fairness perception. These findings suggest that differences in fairness evaluations are driven primarily by the underlying level of disparity, not by the fairness metric itself.

4.4 Participant's beliefs

We conclude this section by looking at participants' beliefs about the choices of others. Overall, participants correctly anticipate the choices of fairness metrics and fairness scores of others. In the baseline, 43% of participants correctly guessed *accuracy* as the most frequently chosen fairness metrics; 57% arrived at the wrong predictions: 26% believed it was Statistical Parity; 24% believed it was Equality of Opportunity, and finally 7% believed it was True Positives Comparison. The comparison of the guessed and true distribution of chosen fairness metrics reveals no difference ($p = 0.41$, exact multinomial test).

Interestingly, around 65% of participants indicate that the fairness metric they had chosen themselves would be the most frequently chosen fairness metric by others.

When asked to predict the average fairness scores, only about 20% of participants in the baseline guessed the correct range for the actual mean general fairness score. 32% overestimated the actual mean general fairness score and about 48% underestimated it.

For both types of beliefs, the guessed fairness metric and the guessed range for the fairness score, we elicited self-confidence of participants in their estimate (from 0 to 100%). The average confidence was 61% in both cases, suggesting that participants were not absolutely sure about the precision of their guesses.

In the treatment, participants were first asked to guess the most frequently chosen utility matrix. 47% correctly predicted utility matrix 2 while 53% made wrong predictions: 38% mistakenly chose utility matrix 3, and 15% mistakenly chose utility matrix 1. The comparison of the guessed and true distribution of chosen utility matrices reveals no difference at 5% ($p = 0.051$, exact multinomial test).

Similarly to the baseline, around 62% of participants indicated the utility matrix they had chosen themselves as the most frequently chosen utility matrix by others. Regarding the deservedness argument, around 53% correctly guessed option 2 (i.e., only comparing utilities of those with true $Y = 1$) to be the most frequently chosen option. Again, around 65% of participants indicated their own chosen argument as the one most likely picked by the others.

Similarly to the baseline, 25% of participants correctly indicated the range of the true average fairness score; 62% underestimated true fairness perceptions, and 13% overestimated it. Finally, participants' average self-confidence were 66% and 56% (on 0–100% scale) for the beliefs about the selected utility matrix and the average fairness score, respectively, suggesting that participants were not different from the baseline in the confidence of their estimates.

We therefore conclude that the two experimental conditions were similar in terms of participants' beliefs. In both conditions, the most common strategy was to submit "naive" beliefs, i.e., to predict one's own choice as the most frequent choice of others.

5 Discussion

This experiment provides important theoretical insights into how the structured exposure to a fairness evaluation framework, which prompts people to consider the utility of the relevant groups and deservedness arguments, can influence preferences for fairness metrics and overall fairness perceptions of resulting algorithmic decisions. Although exposure to such a utility-based framework led to a significant shift in metric preferences, with more participants relying on comparing the average utilities across the groups, there were no significant differences in overall fairness perceptions across conditions.

Prompting participants to explicitly accommodate the utility of the affected individuals and deservedness arguments represents a novel contribution to the empirical literature on perceptions of fairness metrics, encouraging participants to critically assess both fairness outcomes and the moral justifications behind resource allocation decisions. This suggests that fairness metric preferences are

shaped not only by equity considerations but also by deeper moral evaluations of deservedness. Importantly, only a minority of participants apply the fairness metric "true positives comparison" in the baseline. Recall that this metric is, by design, the most consistent when comparing the average utilities across the demographic groups. This suggests that even though all the relevant information was provided (in the form of the confusion matrices), participants rarely rely on comparing the utilities themselves without explicit instruction. In fact, participants in the treatment condition, who were guided in their choice of fairness metric and had to consider utility explicitly, were more likely to choose this metric.

We tested for one particular type of deservedness through cost-efficiency. In our scenario, allocating people with true $Y = 1$ to the job coaching program assured that the coaching resources were spent efficiently. Interestingly, only about half of the participants agreed with this argument in this setting. The other half considered all job seekers to be equally deserving of participation in the job coaching program, highlighting again the context-dependency and individual differences of fairness perceptions. Despite focusing on a specific deservedness argument, we believe that our results might extend towards other scenarios where true labels serve as justifiers for differentiated treatment [26].

Importantly, our results demonstrate that people's perceptions of fairness are malleable: by providing specific types of instructions or guidance, it is possible to shift participants' preferences. Globally, such guidance or "nudging" can be employed for both benign and malicious purposes. In this paper, we have shown that the guidance used is socially beneficial, as it enables participants to consider multiple factors in their fairness judgments and thereby make more informed decisions. However, the potential for malicious use poses a significant risk of manipulating people's preferences in a particular direction. Therefore, the design of any guidance related to fairness evaluations must be approached with caution: if the information provided appears biased or misleading, regulatory oversight may be necessary to ensure that it is revised and that such risks are mitigated.

6 Conclusion

In conclusion, this study demonstrates the potential of utility-based frameworks to alter individuals' preferences for fairness metrics, with a significant shift toward considering the average utilities for the two demographic groups. However, fairness perceptions remained relatively stable across conditions. What ultimately mattered for perceived fairness was the degree of disparity between the considered groups produced by a particular fairness metric. Since fairness metrics generally produce different disparities among the groups, using any particular evaluation framework that "funnels" participants towards specific fairness metrics will indirectly affect their fairness judgments. Future research could explore the long-term effects of such frameworks, whether similar effects occur in other contexts, such as healthcare or criminal justice, and for other socially sensitive attributes (e.g., gender, ethnicity) and whether other types of deservedness arguments (e.g., merit, need, or effort) have similar effects on people's preferences over fairness metrics and resulting fairness evaluation of algorithmic decisions.

7 Limitations

We conclude the paper by acknowledging several important limitations that influence the interpretation of our results. First, our analysis of fairness perceptions focused exclusively on race as the sensitive attribute. Consequently, we cannot determine whether these findings generalize to other prominent social categories such as gender or ethnicity. For the same reason, we are unable to extrapolate our results to more complex, real-world contexts involving intersectionality–that is, situations in which individuals simultaneously belong to multiple social groups, potentially amplifying the sensitivity of fairness considerations. Second, our data on fairness perceptions were collected from a sample of STEM students, who evaluated the scenario as external observers rather than as individuals directly affected by algorithmic decisions. Moreover, the scenario described a situation taking place in a different country, which may have further increased the psychological distance from the problem at hand. These factors limit the generalizability of our findings to the broader population, where people might experience such decisions personally or within a closer sociocultural context (e.g., involving their own country or community). Third, our study relies on hypothetical scenarios and self-reported evaluations rather than real-world decisions or observed behavior. While this approach allows careful experimental control, it may not fully capture how participants would respond when personally affected or under real stakes, potentially limiting ecological validity. Fourth, although we augmented our reporting with effect sizes, some conditions had modest sample sizes, which may reduce the precision of certain estimates and increase the risk of Type II errors for comparisons that did not reach statistical significance. Finally, our measures of fairness perceptions and utility evaluations are task-specific and scenario-dependent. The observed preferences for fairness metrics, utility matrices, and deservedness arguments may

differ in other contexts or with alternative framings, highlighting that the findings should be interpreted as indicative rather than definitive.

Author contributions All authors contributed equally to the conception, design, execution, and writing of this study. All authors reviewed and approved the final manuscript.

Funding Open access funding provided by University of Zurich. The funding from the grant number 187473 from the Swiss National Science Foundation is greatly appreciated.

Data availability The data and the replication package are available here: https://doi.org/10.17605/OSF.IO/SP2VU.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval We thoroughly followed our university's ethical guidelines for IRB approval. First, we obtained informed consent from the participants after they read the experimental instructions. Second, participants were instructed that their participation was voluntary and that they could withdraw from the study at any time. Based on our university guidelines, no further IRB approval was required, as our study did not impose any risks for participants or involve any forms of deception.

References

1. Rabonato, R.T., Berton, L.: A systematic review of fairness in machine learning. AI Ethics **5**, 1943–1954 (2025). https://doi.org/10.1007/s43681-024-00577-5
2. Shams, R.A., Zowghi, D., Bano, M.: Ai and the quest for diversity and inclusion: a systematic literature review. AI Ethics **5**, 411–438 (2025). https://doi.org/10.1007/s43681-023-00362-w
3. Newman, D.T., Fast, N.J., Harmon, D.J.: When eliminating bias isn't fair: algorithmic reductionism and procedural justice in human resource decisions. Organ. Behav. Hum. Decis. Process. **160**, 149–167 (2020). https://doi.org/10.1016/j.obhdp.2020.03.008
4. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. Big Data Soc. **3**(2), 2053951716679679 (2016). https://doi.org/10.1177/2053951716679679
5. Vanderelst, D., Winfield, A.: An architecture for ethical robots inspired by the simulation theory of cognition. Cogn. Syst. Res. **48**, 56–66 (2018). https://doi.org/10.1016/j.cogsys.2017.04.002. (**Cognitive Architectures for Artificial Minds**)
6. Cossette-Lefebvre, H., Maclure, J.: Ai's fairness problem: understanding wrongful discrimination in the context of automated decision-making. AI Ethics **3**, 1255–1269 (2023). https://doi.org/10.1007/s43681-022-00233-w
7. Garcia, A.C.B., Garcia, M.G.P., Rigobon, R.: Algorithmic discrimination in the credit domain: what do we know about it? AI Soc. **39**, 2059–2098 (2024). https://doi.org/10.1007/s00146-023-01676-3
8. Carey, A.N., Wu, X.: The statistical fairness field guide: perspectives from social and formal sciences. AI Ethics **3**, 1–23 (2023). https://doi.org/10.1007/s43681-022-00183-3
9. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al.: AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev. **63**(4/5), 4–1415 (2019)
10. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K.: Fairlearn: a toolkit for assessing and improving fairness in ai. Technical report, Technical Report MSR-TR-2020-32, Microsoft, May 2020. (2020)
11. Corinna, H., Joachim, B., M, L., Christop, H. EWAF, Proceedings of 2nd the European Workshop on Algorithmic Fairness (2023). https://api.semanticscholar.org/CorpusID:259977134
12. Joachim, B., Corinna, H., Michele, L., Christoph, H.: Distributive justice as the foundational premise of fair. ML: Unification, Extension, and Interpretation of Group Fairness Metrics (2023). arXiv:2206.02897. https://doi.org/10.48550/arXiv.2206.02897
13. Starke, C., Baleis, J., Keller, B., Marcinkowski, F.: Fairness perceptions of algorithmic decision-making: a systematic review of the empirical literature. Big Data & Society **9**(2) (2022) https://doi.org/10.1177/20539517221115189
14. Grgić-Hlača, N., Lima, G., Weller, A., Redmiles, E.M.: Dimensions of diversity in human perceptions of algorithmic fairness. In: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. EAAMO '22. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3551624.3555306
15. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), vol. 67, pp. 43–14323. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2017). https://doi.org/10.4230/LIPIcs.ITCS.2017.43
16. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017). https://doi.org/10.1089/big.2016.0047
17. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning: limitations and opportunities. MIT Press, Cambridge, MA (2023). https://fairmlbook.org/
18. Jacobs, A.Z., Wallach, H.: Measurement and fairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, pp. 375–385. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445901
19. Richardson, B., Garcia-Gathright, J., Way, S.F., Thom, J., Cramer, H.: Towards fairness in practice: a practitioner-oriented rubric for evaluating fair ml toolkits. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–13 (2021)
20. Deng, W.H., Nagireddy, M., Lee, M.S.A., Singh, J., Wu, Z.S., Holstein, K., Zhu, H.: Exploring how machine learning practitioners (try to) use fairness toolkits. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 473–484 (2022)

21. Lee, M.S.A., Singh, J.: The landscape and gaps in open source fairness toolkits. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–13 (2021)
22. Srivastava, M., Heidari, H., Krause, A.: Mathematical notions vs human perception of fairness: a descriptive approach to fairness for machine learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19, pp. 2459–2468. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330664
23. Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., Ur, B.: An empirical study on the perceived fairness of realistic, imperfect machine learning models. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20, pp. 392–402. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372831
24. Sengewald, J., Schlichter, A., Siepermann, M., Lackes, R.: Human perceptions of fairness: A survey experiment. In: Beverungen, D., Lehrer, C., Trier, M. (eds.) Conceptualizing Digital Responsibility for the Information Age, pp. 53–70. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-80119-8_4
25. Heidari, H., Loi, M., Gummadi, K.P., Krause, A.: A moral framework for understanding fair ml through economic models of equality of opportunity. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19, pp. 181–190. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3287560.3287584
26. Loi, M., Herlitz, A., Heidari, H.: Fair equality of chances for prediction-based decisions. Econ. Philos. **40**(3), 557–580 (2024). https://doi.org/10.1017/S0266267123000342
27. Feldman, F.: Desert: reconsideration of some received wisdom. Mind **104**(413), 63–77 (1995). https://doi.org/10.1093/mind/104.413.63
28. Miller, D.: Principles of social justice. Harvard University Press, Cambridge, MA (1999). https://doi.org/10.2307/j.ctv1pdrq04
29. Konow, J.: Which is the fairest one of all? a positive analysis of justice theories. J. Econ. Literature **41**(4), 1188–1239 (2003). https://doi.org/10.1257/002205103771800013
30. Tinghög, G., Andersson, D., Västfjäll, D.: Are individuals luck egalitarians? – an experiment on the influence of brute and option luck on social preferences. Frontiers Psychol **8** (2017) https://doi.org/10.3389/fpsyg.2017.00460
31. Cappelen, A.W., Moene, K.O., Skjelbred, S.-E., Tungodden, B.: The merit primacy effect. Econ. J. **133**(651), 951–970 (2023). https://doi.org/10.1093/ej/ueac082
32. Selten, R., Ockenfels, A.: An experimental solidarity game. J. Econ. Behav. Organ. **34**(4), 517–539 (1998). https://doi.org/10.1016/S0167-2681(97)00107-8
33. Lee, M.K., Kusbit, D., Kahng, A., Kim, J.T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., et al. Webuildai: Participatory framework for algorithmic governance. In: Proceedings of the ACM on human-computer interaction 3(CSCW), 1–35 (2019) https://doi.org/10.1145/3359283
34. Mollerstrom, J., Reme, B.-A., Sorensen, E.O.: Luck, choice and responsibility – an experimental study of fairness views. J. Public Econ. **131**, 33–40 (2015). https://doi.org/10.1016/j.jpubeco.2015.08.010
35. Binns, R.: Fairness in machine learning: lessons from political philosophy. In: Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, 149–159 (2018)
36. Charness, G., Rabin, M.: Understanding social preferences with simple tests. Q. J. Econ. **117**(3), 817–869 (2002). https://doi.org/10.1162/003355302760193904
37. Hertweck, C., Heitz, C., Loi. M.: What's distributive justice got to do with it? Rethinking algorithmic fairness from a perspective of approximate justice. Proc. AAAI/ACM Conf. AI Ethics Soc. **7**, 597–608 (2024). https://doi.org/10.1609/aies.v7i1.31661
38. Alkhathlan, M., Cachel, K., Shrestha, H., Harrison, L., Rundensteiner, E.: Balancing act: evaluating people's perceptions of fair ranking metrics. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. FAccT '24, pp. 1940–1970. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3630106.3659018