

Spotting (and Missing) Algorithmic Bias: Investigating User Understanding in a Fairness Assessment Tool

Anna Verheyden^{1,2}, Yizhe Zhang^{1,2}, Robin De Croon^{1,2}, Simone Stumpf^{*3}, Katrien Verbert^{*1,2},

¹ KU Leuven, Department of Computer Science, B-3001 Leuven, Belgium,

² Augment, imec research group at KU Leuven, Leuven, Belgium

³ University of Glasgow, Glasgow, Scotland, UK

{anna.verheyden, yizhe.zhang, robin.decroon, katrien.verbert}@kuleuven.be, simone.stumpf@glasgow.ac.uk

Abstract

Fairness metric selection is typically left to data scientists, but which biases are problematic and which metric captures them best depends on stakeholders' experience and domain knowledge. This calls for involving non-technical stakeholders, but the research prototypes built for this purpose so far have not tested whether these stakeholders form accurate mental models of the metrics they interact with or can act on them to identify biases. We present FairAware, a fairness assessment tool co-designed with Human Resources (HR) domain experts. We evaluate stakeholders' understanding through a mixed-methods study with 70 participants (35 HR employees, 35 job seekers), measuring objective and subjective understanding, cognitive load, bias identification accuracy, and open-ended feedback. Most participants correctly identified the most disadvantaged group, with task duration being the only significant predictor. We also found a gap between subjective and objective understanding, with both groups performing similarly across all measures. These results suggest that fairness assessment tools for non-experts are usable for identifying biases but need built-in checks on understanding before stakeholders make higher-stakes decisions.

Introduction

Recommendation algorithms are increasingly used to support high-stakes decisions in domains such as hiring (Barocas, Hardt, and Narayanan 2019), which strongly affect people's lives. In the hiring domain, algorithmic tools are used to screen resumes, rank candidates, and recommend job matches (Köchling and Wehner 2020). However, this raises concerns around algorithmic bias: if the data or model encodes historical patterns of discrimination, the algorithm may systematically disadvantage certain demographic groups (Barocas and Selbst 2016). Hence, there are ethical and legal reasons to ensure that they are as fair as possible. The EU AI Act (European Parliament and Council of the European Union 2024) now requires human oversight of high-risk AI systems, which means companies are forced to pay attention to assessing the fairness of these systems.

Current ways of assessing fairness rely on fairness metrics. More than fifty metrics have been proposed, many of

which are mutually incompatible (Verma and Rubin 2018; Kleinberg, Mullainathan, and Raghavan 2017), and there is no single correct way to assess algorithmic fairness. Selecting which metrics to apply is itself a value judgment that depends on domain context (Cheng et al. 2021; Lee et al. 2019). Yet, the stakeholders that are best positioned to make that judgment due to their domain expertise or lived experience—HR employees and job seekers—usually lack the technical background to interpret the available metrics.

This raises the question of whether interfaces designed to present fairness metrics to non-technical stakeholders support them in interpreting these metrics correctly and identifying potential algorithmic biases. While recent work has developed fairness tools for eliciting non-expert users' fairness preferences (Cheng et al. 2021; Nakao et al. 2023; Luo et al. 2025), or proposed participatory frameworks that include stakeholders in fairness-related decisions (Cheng et al. 2021; Lee et al. 2019), the question of these stakeholders' understanding of it or their ability to find algorithmic biases has not been investigated.

To fill this gap, we present FairAware, a fairness assessment dashboard to evaluate recommendations for matching job vacancies, co-designed with non-technical stakeholders in the hiring domain, and a mixed-methods study with 35 HR employees and 35 job seekers exploring two research questions:

- (RQ1) What affects users' subjective and objective understanding of fairness metrics, and their ability to identify biases?
- (RQ2) How do users employ the dashboard to advance their understanding and find biases?

This work contributes:

- an artifact: FairAware, a fairness assessment dashboard co-designed with HR domain experts that lets non-technical users navigate four fairness metrics, weigh them according to their preferences, and drill down to details or a causal view for attribute relations.
- empirical findings: A mixed-methods study with 70 non-technical participants showed that most could correctly identify the most disadvantaged group, with time on task as the only significant predictor. Using a new *mental model* instrument we constructed, we found that subjective understanding tracked prior familiarity rather than

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

objective understanding, and dashboard interaction logs did not predict either, suggesting a misalignment between subjective and objective understanding and that participants spotted biases without necessarily understanding the underlying metrics.

Background and Related Work

Measuring Fairness and Involving Stakeholders to Assess Fairness

There is no single, overarching metric for algorithmic fairness. Instead, over fifty metrics have been proposed, each one reflecting a different ethical viewpoint (Verma and Rubin 2018). These fairness metrics are mutually incompatible (Kleinberg, Mullainathan, and Raghavan 2017), making metric selection a context-dependent value judgment (Waller, Rodrigues, and Cocarascu 2025; Luo et al. 2026). Group-level metrics (e.g., statistical parity, disparate impact) compare outcomes between a privileged and an unprivileged group defined by a sensitive attribute, while individual-level metrics (e.g., consistency, counterfactual fairness) ask whether similar people are treated similarly (Verma and Rubin 2018; Kearns et al. 2018). Applying any of these metrics requires two choices that are usually made by data scientists rather than the affected stakeholders: which metric to choose, and which attribute defines the groups. Most existing toolkits operationalize this for data scientists, for example, Fairlearn (Bird et al. 2020), Aequitas (Saleiro et al. 2019), AI Fairness 360 (Bellamy et al. 2019), the What-If Tool (Wexler et al. 2020), or offer a specialized functionality, for example, FairVis (Cabrera et al. 2019) addressing intersectional bias, or Silva (Yan et al. 2020) and D-BIAS (Ghai and Mueller 2023) visualizing causal relationships between attributes.

Because fairness is context-dependent and value-laden, recent research has started to involve non-technical stakeholders, either by eliciting their fairness notions without pre-defining the metrics (Nakao et al. 2022), or by surfacing their varying preferences once metrics are explained (Cheng et al. 2021; Lee et al. 2019; Luo et al. 2026). Several research prototypes have been built to support non-technical stakeholders in assessing fairness: FairHIL (Nakao et al. 2023), the EARN Fairness Framework with the FEE tool (Luo et al. 2025), and the Fairness Elicitation Interface (Cheng et al. 2021), which focus on surfacing harms, eliciting preferences, or facilitating collective deliberation. However, these prototypes assume that stakeholders understand the metrics they interact with, an assumption that has not been tested. If their comprehension is inaccurate, the value of their input is undermined. They have also been evaluated on loan application (FairHIL, FEE) or child welfare prediction (FEI), but not in recruitment, where stakeholder needs and fairness preferences might differ.

Our work takes inspiration from the above studies by building an interface that targets non-technical stakeholders, but differs from these approaches in two ways: (1) we focus on measuring whether non-technical stakeholders actually understand, and can act on, the fairness metrics and the tool presented to them, and (2) we do so in the recruitment

domain, a domain where the EU AI Act applies directly. For (1), we validate their objective understanding, measure whether it aligns with their subjective understanding, and assess its relation to cognitive load. We involve stakeholders throughout the full process, which introduces specific stakeholder needs.

Measuring Understanding

Users' subjective and objective understanding of AI systems and tools is frequently assessed in eXplainable AI (XAI). Hoffman et al. provides a comprehensive overview of methods to do so, including ways to measure users' mental models. A number of question types have been developed to elicit and measure mental models (Kulesza et al. 2012; Lakkaraju, Bach, and Leskovec 2016; Bhattacharya et al. 2024, 2025; Nimmo et al. 2024): 1) fact questions that test retention of specific terminology, 2) simulation or prediction questions that ask users to anticipate system behavior, and 3) articulation questions that require free-text explanations. Often, multiple-choice answers are used in response to these questions, but this format risks overestimating objective understanding because participants can recognize an answer without being able to reproduce it, or they can simply guess (Schuwirth, van der Vleuten, and Donkers 1996).

No validated instruments exist for measuring mental models of fairness metrics specifically. Existing questionnaires target model explanations (e.g., feature attributions, decision rules), vary widely in scoring and item composition, and are very context-specific; no shared guidelines for mental model questionnaire construction exist to our knowledge. Our work contributes the construction process of a new instrument, balancing fact, simulation, and articulation questions to capture understanding at multiple depths, based on prior mental-model studies in XAI. The design choices, scoring, and trade-offs are described in the Methods section.

Methods

Following a human-centered co-design approach (n=15), we first designed a prototype, the FairAware dashboard, to support non-technical stakeholders in assessing fairness and identifying biases. We then conducted a mixed-methods online study with 70 participants (35 HR employees and 35 job seekers) to investigate our main research questions. All research received ethical approval from the KU Leuven Ethics Committee with an approval number G-2025-9565-R2(AMD), and all participants provided informed consent.

Prototype: The FairAware Dashboard

Data and recommendation algorithm. FairAware visualizes the fairness of a recommendation algorithm that suggests matching vacancies to job seekers. The underlying dataset is synthetic and was generated for this study (more information on dataset construction is available in Appendix A (Verheyden et al. 2026)). We constructed 10000 candidate records with seven attributes selected together with our HR partner as the most relevant for fairness assessments in this domain: age, gender, education level, nationality, work experience, employment gap, and target-language fluency. Common biases (for age, gender, and language fluency) were

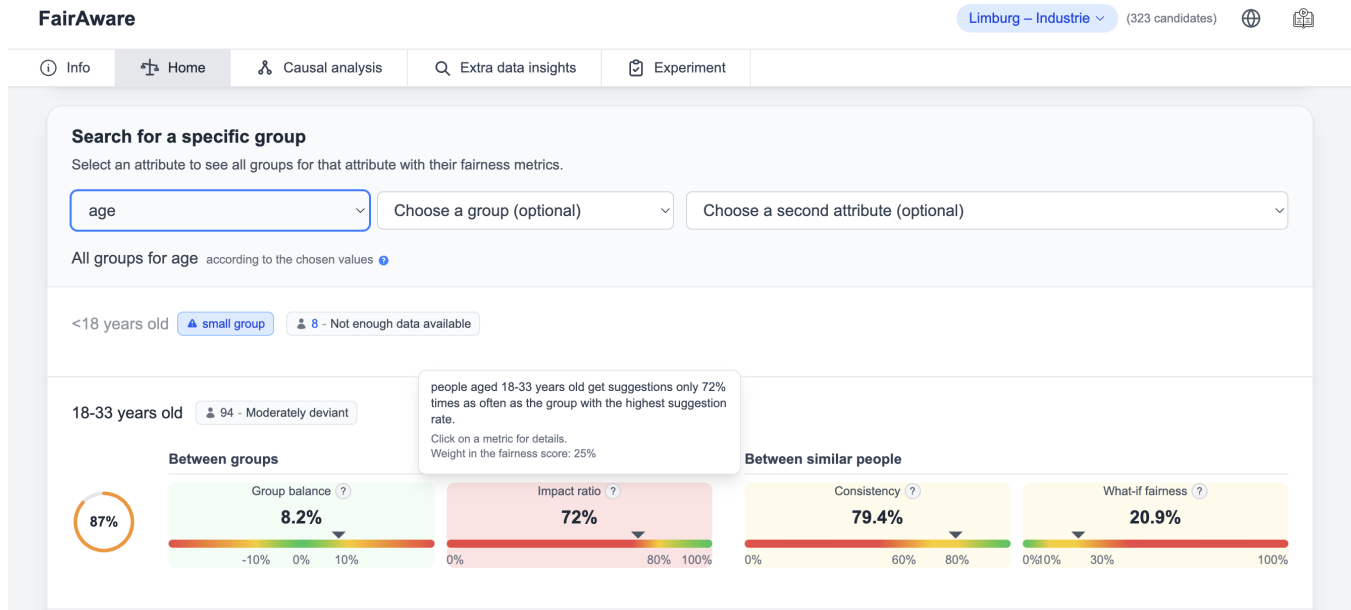


Figure 1: The *Home Page* of FairAware, more specifically, the component to search the results for a specific attribute. Here, age is chosen, and the results are visible for the group 18-33 years old. The 87% on the left is the ‘overall fairness score’, which is the weighted sum of the four metrics to the right of it (after rescaling), based on the participant’s chosen metric priorities.

incorporated based on interviews with HR employees, enabling us to verify the results. Vacancies were grouped by region and sector following the local public employment service’s job-cluster taxonomy (VDAB n.d.), as requested in user interviews. The recommendation algorithm matches candidates to vacancies on a skills similarity score and returns a binary “suggest this vacancy/do not suggest” output per pair, which is what the four fairness metrics in FairAware are computed over. It is therefore a job-matching recommender that shows vacancies to candidates, not a résumé screener or a candidate ranker: bias manifests as unequal access to vacancy suggestions.

The FairAware dashboard. FairAware (Figure 1) is the interface we built to present the fairness of the recommendation algorithm described above (or any other algorithm) to non-technical users. It was co-designed with HR domain experts through biweekly meetings with an industrial partner with over 15000 employees, semi-structured interviews with inclusion experts (n=4), a feedback session with the company’s innovation board (n=6), and five semi-structured interviews with HR employees and a job seeker (n=5). Transcripts were thematically coded following a hybrid inductive-deductive approach (Braun and Clarke 2006). From the resulting codebook (Appendix E (Verheyden et al. 2026)), we identified three main themes, which we translated into the following design goals:

(DG1) Manage cognitive load, operationalized by letting users drill down from the overall score to metric-specific graphs, separating the *Home*, *Experiment*, and *Causal Analysis Page*, and hover-cards with one-sentence metric explanations.

(DG2) Protect candidate privacy by showing only aggregated group-level data and warning when a group is too small to show significant results.

(DG3) Support stakeholder agency, as participants repeatedly asked to run their own queries in addition to the pre-selected results. This is implemented through user-defined attribute binning (e.g., customizable age ranges or employment gap thresholds), free attribute selection on the home tab, and global or per-attribute weight sliders (with optional preset profiles), to configure their own metric preferences.

The dashboard presents biases according to four fairness metrics, two at the group level and two at the individual level, selected based on prior literature (Taka et al. 2024; Lee et al. 2019; Kearns et al. 2018; Verma and Rubin 2018; Luo et al. 2025; Waller, Rodrigues, and Cocarascu 2025). For each metric, the participant-facing label is followed by the formal name in parentheses.

Group balance (statistical parity) compares the vacancy suggestion rate of a given group to the average vacancy suggestion rate. A value of zero indicates parity. Chosen for its alignment with human intuition (Cheng et al. 2021).

Impact ratio (disparate impact) is the suggestion rate of a group divided by the rate for the most privileged group. Values below 0.8 are typically flagged as discrimination. Chosen for its familiarity in the organizational context.

Consistency (consistency) measures, for each candidate, what fraction of their k-nearest neighbors in feature space received the same recommendation. The dashboard reports the mean for the group. When the user clicks

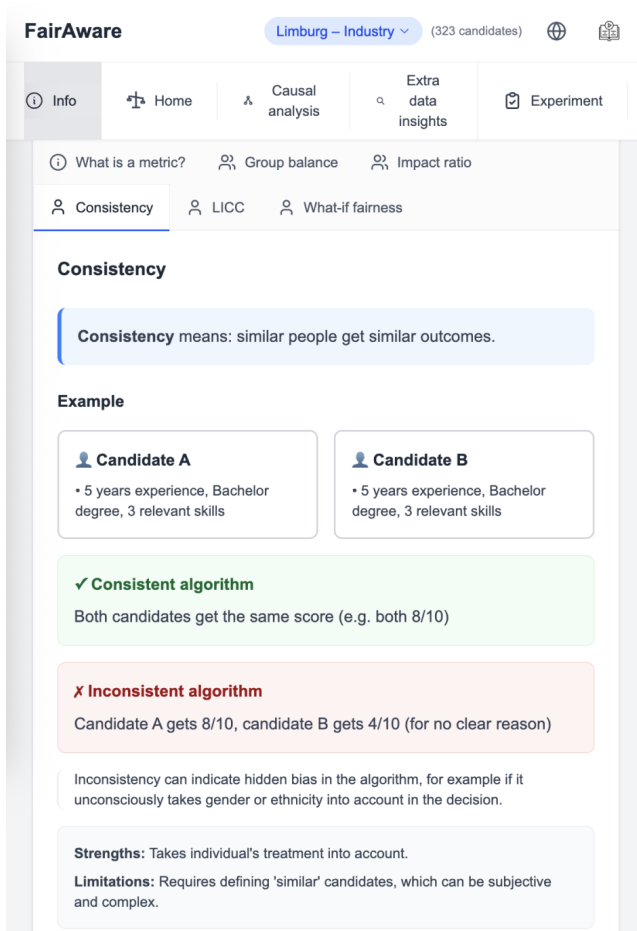


Figure 2: The *Information Page* of FairAware. On top, the different metrics can be selected to see their explanation. Currently, consistency is selected. The one-sentence explanation is shown, followed by an example, extra relevant info, a strength, and a limitation of the metric.

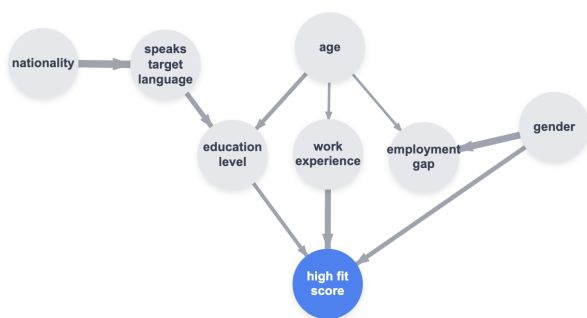


Figure 3: The causal graph of the *Causal Analysis Page* of FairAware. Nodes show the attributes, edges show the correlations, with edge thickness indicating the strength of the correlation. High fit score indicates the similarity score between a vacancy and a candidate, which is used to decide whether they get a vacancy suggested.

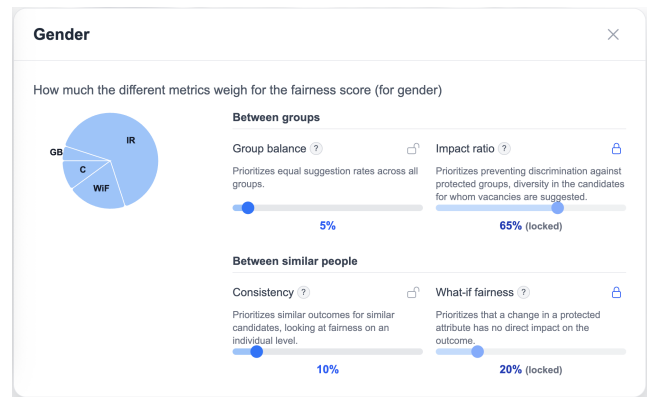


Figure 4: The weight component of the *Experiment Page*. Metrics' impact on the overall fairness score can be adapted by changing the weight for that metric. This can be done using the sliders or presets (not visible here). This shows the component for gender, but it can be adapted either globally or for any specific attribute.

on a result, it is complemented with Low-Individual-Consistency-Candidates (LICC, the absolute count of inconsistently treated people in the group (Waller, Rodrigues, and Cocarascu 2025)) in the more detailed view.

What-if fairness (counterfactual fairness) asks 'if a candidate's sensitive attribute were flipped (e.g., female to male), would the recommendation change?'. The metric reports the proportion of candidates for whom the recommendation would flip.

There are five tabs, each with its own functionality available to participants:

- The *Home Page* (Figure 1): the main page of the dashboard, where users can see the fairness for the three different demographic groups that experience the highest bias, according to an "overall fairness score". Below it, users can select any attribute (e.g., age) to see fairness results for specific groups, or select an intersection of two attributes (e.g., 'women aged 18-33'). For each group, the fairness is shown as the scores of the four fairness metrics, as well as an overall 'fairness score'. This fairness score is computed by rescaling the four metrics to a shared scale that represents how 'good' the value is for that metric, and then computing the weighted sum of these four values, based on the metric priorities that the user has chosen. For every metric value, hovering shows a one-sentence natural-language explanation, and clicking opens a detailed graph (e.g., for consistency, a bar chart showing what fraction of similar candidates received the same recommendation per group).
- The *Information Page* (Figure 2): the start page upon first use, containing explanation cards for each of the four metrics, with a one-sentence definition, a short explanation, and an example. This tab does not change with user input.
- The *Experiment Page*: lets the user 1) re-bin attributes (e.g., adapt the age ranges or the number of months that

defines an employment gap) and 2) re-weight the four metrics according to their preferences (Figure 4), to configure their own fairness policy. Changing weights immediately updates the overall fairness score on the *Home Page*.

- The *Causal Analysis Page*: contains a directed graph (Figure 3) over the seven attributes and the ‘fit score’ (the similarity score of how well the candidates’ skills match a vacancy, used to compute the binary match). Edge thickness encodes the strength of the relationship. Users have the option to give feedback on edges that they find unacceptable, but this does not change anything on the other pages. This page can be used when a user finds a bias and wants to know what other attributes might have influenced this (e.g., if we want to solve the work experience bias, should we also look at gender?).
- The *Extra Data Insights Page*: contains data distribution histograms and a per-metric ranking of biased groups. This provides more information on the dataset.

We conducted a face-to-face pilot with five participants from an HR company (two recruiters, two inclusion-focused HR employees, one job seeker) prior to the study. This surfaced five design issues: confusion about the way that metrics were weighted, unnoticed small-group warnings, the recurring suggestion to separate exploration from the home page, a misinterpreted metric calculation from the example, and low visibility of the gradient bars that indicate a metric value’s severity. This led to five changes in the final tool: a more explicit explanation of the weighted ‘overall fairness score’, a split to separate the *Home Page* and the *Experiment Page*, traffic-light background colors added to the gradient bar, hidden results for small groups (only name, size, and ‘not enough data’ label shown), and using unequal group sizes in the metric examples.

Study Setup

Participants. We recruited 70 participants via Prolific, a platform to recruit online participants in studies (Prolific 2026). This sample included 35 HR professionals (24 women and 11 men, aged 21-57, $M=33.6$, $SD=8.5$) and 35 job seekers (11 women and 24 men, aged 18-58, $M=26.6$, $SD=8.2$). Both groups were included as they represent two sides of algorithmic hiring with different stakes.

Participants had to be fluent in English, with at least 80% task approval ratings, and aged 18+. Prolific screeners were used, complemented with our own screener questions (Appendix B (Verheyden et al. 2026)) to validate appropriate HR expertise. Two participants were excluded and replaced for quality reasons. Participants were compensated at £9/hr.

Procedure. The study was conducted online and lasted around one hour ($M=70.89$ minutes). After the screening, it followed three phases (Figure 5). In Phase 1, participants gave informed consent, went through a structured dashboard tutorial (interactive pop-ups that they had to click through, $M=8.92$ min), and completed pre-study questionnaires ($M=25.20$ min) covering demographics, familiarity with fairness metrics, statistics, and interpreting graphs, as

well as objective understanding (OU). In Phase 2 ($M=16.61$ min), they performed three tasks: 1) they identified the most biased group according to an attribute of choice, 2) they configured metric weights to reflect their own fairness policy, and 3) they investigated the cause behind a bias. Finally, in Phase 3, they completed a post-study questionnaire repeating the OU instrument and the NASA-TLX cognitive load questionnaire (Hart and Staveland 1988), a four-question subjective understanding scale adapted from Hoffman et al., and five open-ended feedback questions on what helped them the most/least to understand the content, sources of confusion, accessibility, and any further thoughts ($M=16.61$ min). The full questionnaires are available in the supplementary materials (Appendix C (Verheyden et al. 2026)).

Data collection and analysis. We collected objective understanding, subjective understanding, cognitive load, qualitative responses, and logging data, and analyzed each as follows.

We measured *Objective Understanding (OU) through mental model scores*. As no validated instruments exist for measuring mental models of fairness metrics, we constructed one for FairAware, inspired by previous research (Lakkaraju, Bach, and Leskovec 2016; Bhattacharya et al. 2024, 2025; Kulesza et al. 2010, 2012, 2013, 2015; Nimmo et al. 2024; Jalali et al. 2023; Ribeiro, Singh, and Guestrin 2016; Cheng et al. 2019; Bove et al. 2022; Lim, Dey, and Avrahami 2009; Hase and Bansal 2020). First, the nine items (12.5 points in total) are split across three question types, balanced to reflect increasing depth: fact (16%), simulation (36%), and articulation (48%) questions. None of the question types accounts for more than 50%. 72% of the score touches the weighted fairness score, as that is the central concept of FairAware, and 16% explicitly asks about metric definitions, although understanding the metrics is a prerequisite to answering some of the other questions well. Second, multiple-choice questions had at least four options and an ‘I don’t know’ option to reduce guessing, and six out of nine questions were open-ended to allow for a more accurate estimation of respondents’ mental model. Third, item difficulty was calibrated using a pilot to avoid floor and ceiling effects. Open-ended items were scored with rubrics: two researchers independently scored 10% of responses (weighted Cohen’s $\kappa = 0.728$, substantial agreement (Landis and Koch 1977)) before one researcher coded the rest. The same instrument was used pre- and post-task. Pre-task (median=39.60%, $SD=23.10$) and post-task (median=43.2%, $SD=23.01$) scores showed no ceiling or floor effects. We define *Learning Gain (LG)* as the post-task OU score minus the pre-task OU score. OU and LG were tested with paired t-tests or Wilcoxon signed-rank tests when non-normal, and group differences with independent t-tests or Mann-Whitney U.

Subjective Understanding (SU) was measured post-task on four 7-point Likert items adapted from Hoffman et al.’s (Hoffman et al. 2018, 2023) scale, covering predictability, intelligibility, and utility.

Cognitive load (CL) was measured post-task using the six NASA-TLX items on a 7-point Likert scale. Correlations be-

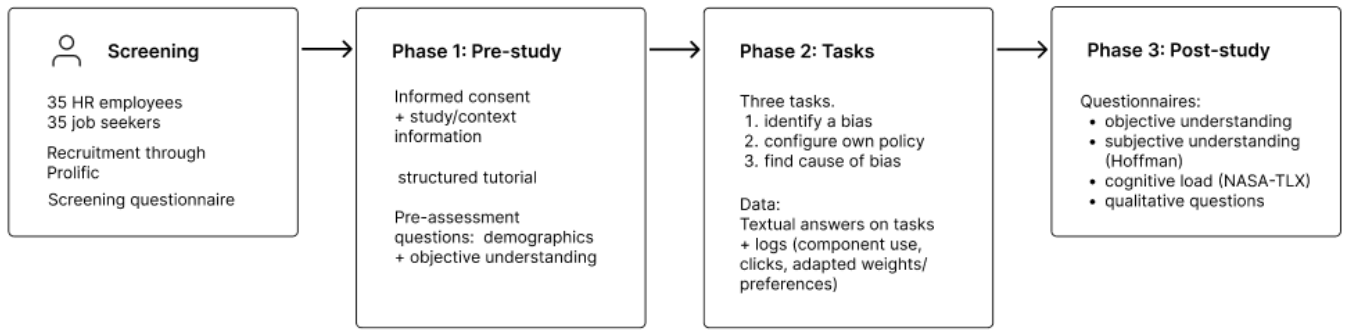


Figure 5: Study setup of the mixed-methods study

tween CL, SU, familiarity, and OU were all measured using Spearman's ρ .

Qualitative responses were coded thematically using a hybrid inductive-deductive thematic analysis (Braun and Clarke 2006) by two researchers on 10% of the data (validated employing unweighted $\kappa = 0.871$) before the remaining coding was done by one researcher. The responses consisted of the five feedback questions and three open-text answers for the three tasks the participants completed. These task responses were filled in while they were performing the tasks. For the first task, they were asked to write down the attribute they found most important in an assessment, and the group that experienced the most bias according to that attribute. The second task asked them to configure their ideal fairness policy and then briefly describe how they approached this. This allowed us to see whether they had clear preferences or priorities, and, if they did not change much, whether that was due to being content with the current configuration or to not understanding how to adapt it. The third task asked them to investigate what other attributes might have influenced the bias identified in the first task. They had to write down one of those attributes and briefly explain how they knew this.

Logging via Microsoft Clarity and our own logs captured time per study phase, hovers, clicks on metric results, tab visits, attribute searches, and parameter changes (the latter was needed to score task 1, since changed weights alter the fairness score). Finally, page blurring and fast returns were tracked as a heuristic for genAI use. Logging results were tested against OU using Spearman's ρ , and between groups with t-tests or Mann-Whitney U. Throughout, statistical significance is denoted as $\dagger p < .10$, $* p < .05$, $** p < .01$, and $*** p < .001$.

Results

We organize the results by research question (RQ). Across all measures and both RQs, job seekers and HR employees did not differ significantly (Figure 6); we therefore report the combined results below.

RQ1: What Aspects Affected Users' Understanding and Ability to Find Biases?

Subjective understanding tracked confidence, not comprehension. The strongest pattern in our data is a divergence between how well participants thought they understood the tool (SU) and how well they actually did (OU). SU correlated only weakly with post-task OU ($\rho = 0.262, p = .028^*$). It correlated strongly with prior familiarity with graphs ($\rho = +0.44, p < .001^{***}$), statistics ($\rho = +0.30, p = .012^*$), and fairness metrics ($\rho = +0.51, p < .001^{***}$), and negatively with cognitive load ($\rho = -0.43, p < .001^{***}$) (Figure 7). In other words, participants felt they understood the tool when it looked familiar, and felt they did not when they were overwhelmed, independently of whether they had actually understood the metrics and overall fairness score. The SU-OU correlation held for job seekers ($\rho = 0.342, p = .044^*$), but was weaker for HR employees ($\rho = 0.295, p = .086^\dagger$).

Looking at the participants at the extremes of the SU-OU gap¹ clarifies this pattern. On the high-SU/low-OU side, P11 and P20 scored 4/5 and 5/5 on the HR pre-screening but often gave superficial OU answers (e.g., answering 'explain why you made these matches' with 'because it is very clear'), showing overconfidence and lower engagement; P64 and P68 gave more elaborate answers, but misunderstood specific metrics, noting a low affinity with numbers and feeling overwhelmed by the amount of information as reasons for a lower OU score. The reverse pattern, low-SU/high-OU, showed self-critical assessments rather than weak performance. P36 and P62 lost points due to hesitancy (P36: "I did not made [sic] them because I am not sure how to"), despite otherwise insightful answers, and P43, P52, and P59 rated their SU moderately while providing good OU responses. They reported time pressure and information density as challenges, rather than confusion.

Familiarity helped subjective understanding, but not objective understanding. Prior familiarity correlated only weakly or not at all with OU (graphs: $\rho = +0.26, p = .032^*$; statistics: $\rho = +0.20, p = .099^\dagger$; fairness metrics $\rho = +0.20, p = .098^\dagger$). Taken together with the SU find-

¹Computed per participant as $\left(\frac{(SU-1)}{6} - \frac{(post-taskOU)}{12.5}\right)$, rescaling both to [0,1] before subtracting.

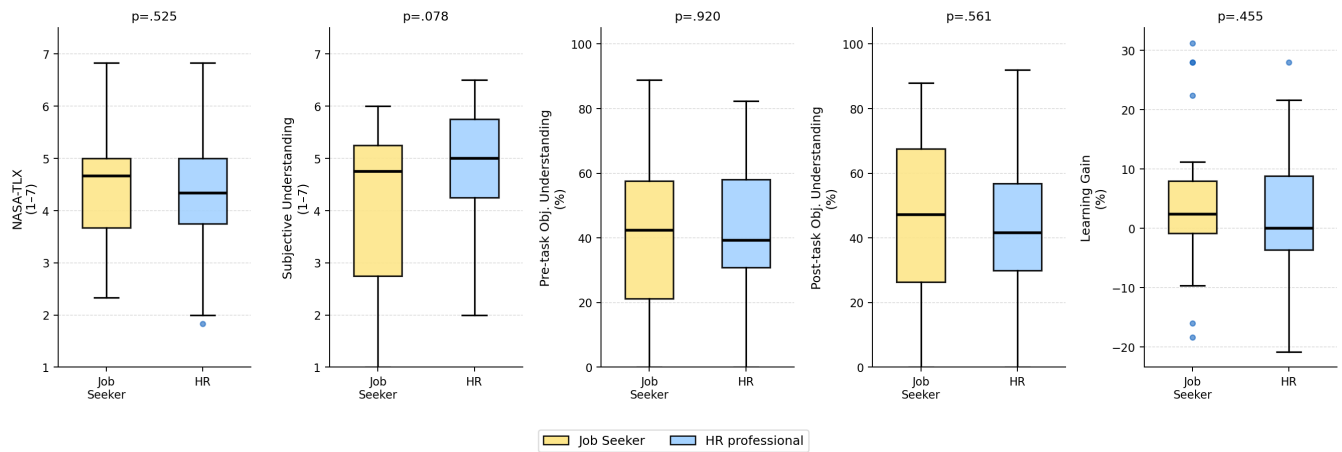


Figure 6: A comparison between job seekers and HR employees on cognitive load, subjective understanding, pre- and post-task objective understanding scores, and the learning gain, which is measured as the change in objective understanding score. None of the differences are significant, implying similar outcomes across stakeholder types.

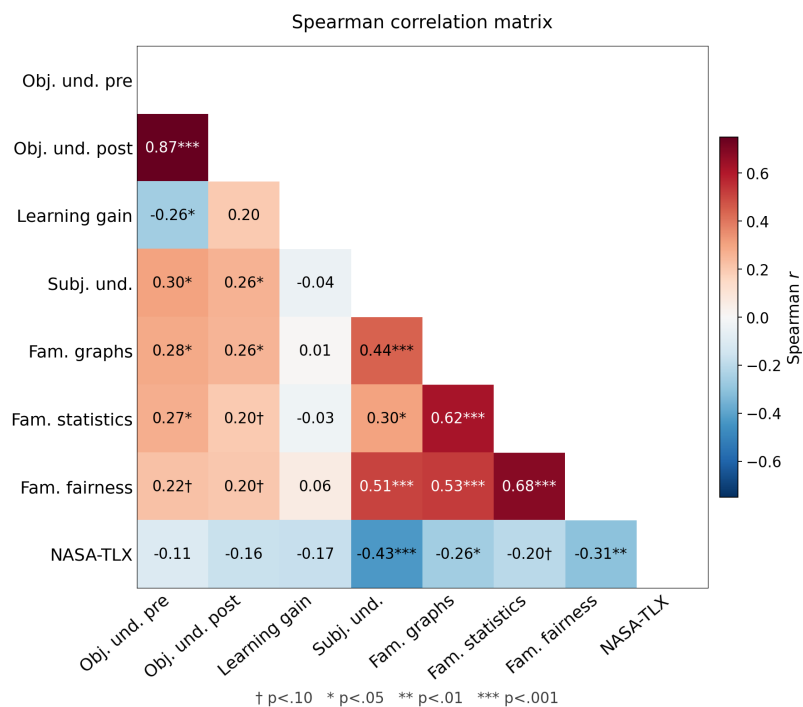


Figure 7: A heatmap of the correlations.

ings above, this shows that familiarity improves SU but not OU. Cognitive Load shows the same pattern, not correlating with OU ($\rho = -0.17, p = .173$) despite the SU correlation.

These two patterns are also connected to a possible confound: that low OU reflects the effort of navigating an interface that presents four metrics and a weighted score at once, rather than the understanding of the metrics themselves. However, if navigation difficulty were the main factor, we would expect it to lower OU alongside SU and an increasing CL, yet CL did not correlate with OU and familiarity improved SU but not OU. This makes a pure navigation-difficulty explanation unlikely, though it does not entirely rule out its contribution.

Most participants were able to spot a bias, but not because of understanding it. Despite the SU-OU divergence, 42 of 70 participants correctly identified the most disadvantaged group for an attribute of choice. Fourteen gave unclear answers, typically naming the attribute ('gender') instead of the group ('female'), and 14 were incorrect (vague answer beyond attribute-group confusion $n=8$, wrong group $n=4$, named a metric $n=1$, 'I don't know' $n=1$). The biggest predictor of success was time on task ($p = .011$): neither pre- nor post-task OU, SU, nor CL predicted who correctly identified the bias.

As for participant characteristics, the 42 participants who identified the correct group differed from the 28 who did not in age (Mann-Whitney $U = 393.5, p = .020$), with successful participants being younger. However, there were no statistical differences due to gender ($\chi^2(1) = 0.54, p = .464$) or stakeholder group ($\chi^2(1) = 0.54, p = .464$). Within HR professionals, years of HR experience was unrelated to success ($\rho = -.152, p = .385$). Aside from a modest age effect, success in finding the bias depended on engagement with the tool, not on who the participant was.

This is consistent with the qualitative pattern that the tool has a learning curve, with early confusion (especially conflating groups with attributes during the first task) that improved with longer interaction. Bias identification after a first short session should therefore not be treated as a decision-ready result; appropriate use of the tool requires more interaction time.

Furthermore, read together with the SU-OU divergence, a non-technical participant can read the interface and point at the worst group, without forming an accurate mental model of why a group is treated worse or what the metrics measure.

Bias-spotting appears to be robust to understanding gaps for the same reason it is robust to expertise levels, as the tool shows which groups score worst at a glance, so the task is decoupled from understanding the underlying metric. A fairness assessment tool can thus be usable for high-level bias detection yet unsafe for higher-stakes participation, because the user might look confident and pick the right group at first sight, but their underlying understanding might be incorrect, meaning that more complex tasks could go wrong too. Any subsequent task that requires the user to trade off metrics, justify a choice, or deliberate with others might depend more on the objective understanding than the bias-spotting accuracy.

RQ2: How Do Users Employ the Dashboard to Advance Their Understanding and Find Biases?

Dashboard use produced a small gain in objective understanding. OU scores varied widely and were low overall (pre-task: median=39.60%, SD=23.10; post-task median=43.20%, SD=23.01). Learning Gain was significant across the full sample (Wilcoxon: $W = 592.0, Z = -2.040, p = .041, r = 0.244^*$), a change driven by job seekers (Wilcoxon: $W = 147.0, Z = -1.979, p = .048, r = 0.335^*$) but not HR employees (paired-t test: $t(34) = 1.000, p = .325, d = 0.169$). The gain is modest in absolute terms but informative given the conditions: Phase 2 lasted 16.61 minutes on average, and the OU instrument included several deliberately challenging items. That a 15-minute first encounter produced a measurable increase in OU suggests that participants' OU would increase over time.

Helpfulness reports showed what logs or dashboard use miss. Qualitative responses identify the *Information Page* as the main page for understanding the content: 28 participants explicitly named it as one of the most helpful components, often noting the need to revisit explanations because of the high amount of new information (e.g., P47: "I had to go back and look at the explanations of what each metric is a few times."). Naming it did not correspond to scoring better on OU ($M=45.0\%$ against a rest-of-sample $M=45.3\%$). The *Information Page* is the page participants default to, but using it does not separate participants who built a better mental model of the metrics from those who did not.

The *Home Page*, named by 22 participants as most helpful, did separate participants on OU: those who named it as most helpful scored on average $M=58.1\%$ against $M=39.3\%$ for the rest of the sample (two-sided Welch's: $p=0.002^{**}$, $d=0.86$, Holm-adjusted $p=0.008^{**}$). Their open responses described their understanding of the overall fairness score and how the different metrics contributed to it. As the *Home Page* summarizes most of the content, participants had to grasp it to find it helpful.

The *Experiment Page* was the most polarizing: 14 participants named it as most helpful, generally as the place where the weighting mechanism started making sense (e.g., P34: "the experiment tab helped me see how changing the weights of the metrics affects the fairness score, [...] easier to understand in a practical way."), and 11 named it as least helpful, generally because they had not yet had time to digest the metrics and the *Home Page* before being asked to manipulate it (e.g., P36: "The experiment. [...] I had too little experience."). An exploratory analysis shows that polarization is not the same across stakeholder groups: 11 of 35 job seekers named the *Experiment Page* as most helpful, against only 3 of 35 HR employees, even though the groups did not differ on OU, SU, or cognitive load. We interpret this as a difference in what the two groups found informative rather than an indication of how much they understood. HR participants might be more used to reading an overview report, whereas a job seeker might put more emphasis on figuring out what policy they should prefer. The response that indicated that a participant did not feel ready to change the

weights suggests that careful sequencing of the task is important. Interactive configuration can be valuable, but only once the static overview is fully internalized.

22 of 70 participants reported some form of overwhelm, substantially more often from job seekers (16/35) than from HR employees (2/35), which we take as another way the experience of the dashboard differs between the groups despite their similar OU, SU, and CL scores.

Self-reported usefulness did not directly translate to actual usage. The interactions with the *Causal Analysis Page* showed some interesting contradictions. Sixteen participants named it as the least helpful, and only ten (five from each group) used it for the task that asked them to identify what other attributes might have influenced a bias, the task for which it was built. Moreover, of the seven participants who named it ‘most helpful’, only two actually used it for that task, and two others could not correctly interpret the graph on the OU questionnaire. This exemplifies the SU-OU gap identified above, now at the component level: participants who could not interpret the graph nevertheless believed that it helped them, and five participants who reported that it was hard to interpret the graph interpreted it correctly (scoring 2/2 on OU).

Exploration depth varied, repeating explanations remains valuable either way. Usage logs did not correlate with OU, SU, or cognitive load. The qualitative responses help explain why. First, exploration depth varied widely. At one end, some participants moved quickly through the tutorial without actively engaging the hover and click features, and when uncertain, they returned to the general explanations in the *Information Page* instead of reading the in-context details. Others fully explored the tool, including the *Causal Analysis Page* and the *Extra Data Insights Page*. Hover and click counts did not serve as strong predictors here, because they conflate different participant strategies: curiously exploring details or being confused and seeking explanations by hovering result in the same behavior. This makes it difficult to distinguish between different ways participants use the available explanations. Second, the participants who moved through the tutorial quickly did not do so because they already understood the tool; the SU-OU pattern shows that several were overconfident. As a result, participants valued the explanations remaining accessible, as they reported in the feedback questions. A takeaway for tool design is that tooltips, tutorials, and hover-explanations should remain visible throughout the usage, not just at the start (e.g., as participants noted: ‘I was confused at first as to what the what-if was comparing until I hovered over it.’, ‘more accessible: Having access to the tutorial’).

Discussion

Implications

Beware of and counteract Mount Stupid. We found that Subjective Understanding rose with familiarity and decreased with Cognitive Load, while Objective Understanding did neither. This aligns with the miscalibration effects documented in the broader psychology literature (Kruger

and Dunning 1999), where people who feel they understand a domain are not necessarily the people who actually do. This raises a design concern for fairness assessment tools: an intuitive-looking or overly simplified interface might lower cognitive load and difficulty of using it, but at the cost of reducing participants’ effort to verify or improve their understanding. Objective Understanding cannot be ensured by simplified UI design; it needs to be checked explicitly as well.

Measuring and reporting both constructs is thus required. Familiarity with fairness metrics was the strongest correlation with SU and was unrelated to OU, so the users most at risk of unwarranted confidence might be the ‘experts’ that end up using the tool. Verification of the level of OU is thus important in any setting. This could be achieved through lightweight, quiz-like prompts embedded in the workflow, similar to those in our mental model questionnaire. In practice, these could be short, auto-scorable checks (e.g., a pop-up with “If this group’s balance worsened, the fairness score would go: up/down/stay the same/I do not know”). Responses could then be used to mitigate areas where understanding is compromised, perhaps through surfacing short explanations. This could identify flawed mental models early on and counteract them. Personalized adjustments could also be made to the interface based on the user’s OU, mirroring work in personalized XAI (Conati et al. 2021).

Dashboard use does not ensure users understand fairness metrics. Hover counts, click counts, and page visits did not predict Objective Understanding, possibly because they do not communicate the user’s goals: the same hover can mean “I am lost” or “I am curious”. This means that help and explanation features should remain visible and accessible from multiple places to provide multiple educational opportunities when they need it. Second, perceived component helpfulness was partially decoupled from actual usage. Of the seven participants who reported the causal graph as most helpful, only two actually used it for the bias-cause task. Similarly, reporting the *Home Page* as most helpful correlated with higher OU. More work is needed to identify truly helpful pages and explanations for understanding fairness metrics.

Algorithmic bias finding is achievable but takes time. Non-technical stakeholders can perform basic fairness assessments with an appropriately designed tool, but the only significant predictor of success was task duration, not OU or SU. However, increasing the time would also likely increase cognitive load, so it is a delicate balance to be struck.

This might necessitate that organizations using fairness assessment tools need to conduct shorter, repeated assessments over time, rather than a longer single session. This would also allow for developing clearer fairness preferences through interaction and learning, and for evolving them, as they are not static (AlRossais 2025).

Consensus-building frameworks need to account for miscalibrated participants. Frameworks like WeBuildAI (Lee et al. 2019) assume that stakeholders bring individ-

ual understanding into group deliberation and that stronger opinions might reflect deeper engagement. However, our results complicate that assumption, because some stakeholders may hold strong opinions based on an incorrect mental model and therefore low Objective Understanding. For example, some people might believe that a high impact ratio means the algorithm is fair when it actually indicates the opposite. This could steer a group discussion towards incorrect conclusions. We argue that consensus-building frameworks thus need to be designed for *calibrated participation*, for example, by gating opinion weighting with a brief understanding check.

Transferring FairAware to other domains. Although FairAware’s content is domain-specific, its structure is generalizable: the four metrics, the rescaled weighted score, the drill-down, and the causal view are defined over any binary decision with a sensitive attribute. The tool could therefore be readily translated to other settings based on tabular data, e.g., loan approval or benefits triage, by replacing the attribute set and the group levels. However, the metrics and group levels have to be based on co-design, since which groups and metrics are meaningful is something that depends on domain context (Cheng et al. 2019; Lee et al. 2019).

Future Work

Three potential future work directions follow from our results. First, longitudinal, real-world studies are needed to see how mental models of non-technical stakeholders evolve with repeated use of fairness assessment tools, whether bias identification accuracy increases, and whether the SU-OU gap decreases or not. Second, our finding that bias-finding accuracy is based on time rather than understanding suggests further research into the effects of more complex tasks, such as policy configuration, metric trade-offs, or justification of assessment decisions on SU, OU, and the ability to assess biases. Third, the consensus-building implication motivates empirical research that investigates how understanding affects deliberations of fairness decisions within groups.

Limitations

Four shortcomings qualify these results. First, some participants completed the tasks quickly, which might limit engagement with the complex concepts we presented in the study. Second, we presented only four fairness metrics in our dashboard. Although these were selected during co-design as the most relevant, this work could be extended to include more fairness notions, some of which could be user-specified (Luo et al. 2026). Third, the study was conducted online via Prolific, in a one-time session with no consequences attached to participants’ judgments, using synthetic data. While we used screening to ensure the quality of responses, an online setting offers less control than an in-person study. Indeed, authentic data might be more difficult to assess but provide a high-stakes setting, which could influence engagement and calibration. More work is warranted to study the use of tools in practice. Fourth, no validated questionnaire exists for measuring mental models of fairness

metrics or assessment tools. Our instrument discriminated well between levels of understanding without ceiling or floor effects, but absolute scores cannot be directly compared to those from other mental-model questionnaires. Standardization of OU measurements is desperately needed.

Conclusion

We tested a common assumption in algorithmic fairness work: that the non-technical stakeholders asked to choose between or interpret fairness metrics actually understand those metrics. To do this, we (1) co-designed FairAware, a fairness-assessment dashboard for non-technical stakeholders in the HR domain, (2) built a mental-model instrument grounded in prior literature to measure their objective understanding, and (3) ran a mixed-methods study with 35 HR employees and 35 job seekers.

The assumption does not always hold. Across 70 participants, subjective understanding correlated with prior familiarity but only weakly with objective understanding; HR employees were not better calibrated than job seekers; and the only predictor of correctly identifying a biased group was time spent, rather than understanding. Three implications follow:

- For tool developers: embed lightweight comprehension checks into the workflow rather than relying on user-reported confidence. These checks can then be used to assess the reliability of these users’ takeaways from the tool, as well as serve to provide short, targeted educational opportunities where miscomprehension occurs frequently.
- For participatory fairness research: when aggregating preferences into a shared policy, weight stakeholder input by calibrated comprehension rather than familiarity or subjective understanding. When conducting group discussions, take the objective understanding into account as well (e.g., by first testing it and providing a short education on miscalibrated aspects).
- For deployment: treat single-session assessments as preliminary. The EU AI Act’s ‘meaningful human oversight’ requirement is not met by a confident user who has quickly read the interface for the first time, but not grasped the metrics.

Our work contributes a stakeholder-facing tool and empirical evidence that comprehension checks are needed before stakeholder input can be trusted. Our mental model instrument offers a starting point that others can adapt to their own tools and metrics. Together, these contributions support the more responsible integration of stakeholders into fairness assessments and, through that, more responsible AI.

Acknowledgments

This work has been funded and supported by the imec.icon CAPTURE project and the Flanders AI Research Program (FAIR), with project support from imec and VLAIO (Flanders Innovation & Entrepreneurship)(Grant No. HBC.2024.0220). Furthermore, we would like to thank all participants for their time and valuable feedback.

References

- AlRossais, N. A. 2025. Time-Evolving Fairness and Accuracy in Recommendation Systems: A Stereotype-Based Framework for Addressing Trade-Offs. *IEEE Access*, 13: 178758–178771.
- Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. MIT Press.
- Barocas, S.; and Selbst, A. D. 2016. Big data’s disparate impact. *California Law Review*, 104(3): 671–732.
- Bellamy, R. K.; Mojsilovic, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; Zhang, Y.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; and Mehta, S. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63.
- Bhattacharya, A.; Stumpf, S.; Croon, R. D.; and Verbert, K. 2025. Explanatory Debiasing: Involving Domain Experts in the Data Generation Process to Mitigate Representation Bias in AI Systems. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. ISBN 9798400713941.
- Bhattacharya, A.; Stumpf, S.; Gosak, L.; Stiglic, G.; and Verbert, K. 2024. EXMOS: Explanatory Model Steering Through Multifaceted Explanations and Data Configurations. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. ISBN 9798400703300.
- Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.
- Bove, C.; Aigrain, J.; Lesot, M. J.; Tijus, C.; and Detyniecki, M. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *International Conference on Intelligent User Interfaces, Proceedings IUI*.
- Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3.
- Cabrera, A. A.; Epperson, W.; Hohman, F.; Kahng, M.; Morgenstern, J.; and Chau, D. H. 2019. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *2019 IEEE Conference on Visual Analytics Science and Technology, VAST 2019 - Proceedings*.
- Cheng, H. F.; Stapleton, L.; Wang, R.; Bullock, P.; Chouldechova, A.; Wu, Z. S.; and Zhu, H. 2021. Soliciting stakeholders’ fairness notions in child maltreatment predictive systems. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. ISBN 9781450380966.
- Cheng, H. F.; Wang, R.; Zhang, Z.; O’Connell, F.; Gray, T.; Harper, F. M.; and Zhu, H. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Conference on Human Factors in Computing Systems - Proceedings*.
- Conati, C.; Barral, O.; Putnam, V.; and Rieger, L. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artif. Intell.*, 298(C).
- European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. Entered into force 1 August 2024.
- Ghai, B.; and Mueller, K. 2023. D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias. *IEEE Transactions on Visualization and Computer Graphics*, 29.
- Hart, S. G.; and Staveland, L. E. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Hancock, P. A.; and Meshkati, N., eds., *Human Mental Workload*, volume 52 of *Advances in Psychology*, 139–183. North-Holland.
- Hase, P.; and Bansal, M. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for Explainable AI: Challenges and Prospects Institute for Human and Machine Cognition. *arXiv preprint arXiv:1812.04608*.
- Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance.
- Jalali, A.; Haslhofer, B.; Kriglstein, S.; and Rauber, A. 2023. Predictability and Comprehensibility in Post-Hoc XAI Methods: A User-Centered Analysis. In *Lecture Notes in Networks and Systems*, volume 711 LNNS.
- Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *35th International Conference on Machine Learning, ICML 2018*, volume 6.
- Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent trade-offs in the fair determination of risk scores. In *Leibniz International Proceedings in Informatics, LIPIcs*, volume 67.
- Kruger, J.; and Dunning, D. 1999. Unskilled and unaware of it: How difficulties in recognizing one’s own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6): 1121–1134.
- Kulesza, T.; Burnett, M.; Wong, W. K.; and Stumpf, S. 2015. Principles of Explanatory Debugging to personalize interactive machine learning. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, volume 2015-January.
- Kulesza, T.; Stumpf, S.; Burnett, M.; and Kwan, I. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Conference on Human Factors in Computing Systems - Proceedings*.

- Kulesza, T.; Stumpf, S.; Burnett, M.; Wong, W. K.; Riche, Y.; Moore, T.; Oberst, I.; Shinsel, A.; and McIntosh, K. 2010. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *Proceedings - 2010 IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC 2010*.
- Kulesza, T.; Stumpf, S.; Burnett, M.; Yang, S.; Kwan, I.; and Wong, W. K. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC*.
- Köchling, A.; and Wehner, M. C. 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.
- Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-August-2016.
- Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.
- Lee, M. K.; Kusbit, D.; Kahng, A.; Kim, J. T.; Yuan, X.; Chan, A.; See, D.; Noothigattu, R.; Lee, S.; Psomas, A.; and Procaccia, A. D. 2019. Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3.
- Lim, B. Y.; Dey, A. K.; and Avrahami, D. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Conference on Human Factors in Computing Systems - Proceedings*.
- Luo, L.; Nakao, Y.; Chollet, M.; Inakoshi, H.; and Stumpf, S. 2025. EARN Fairness: Explaining, Asking, Reviewing, and Negotiating Artificial Intelligence Fairness Metrics Among Stakeholders. *Proceedings of the ACM on Human-Computer Interaction*, 9: 1–37.
- Luo, L.; Nakao, Y.; Chollet, M.; Inakoshi, H.; and Stumpf, S. 2026. "I Think This Is Fair": Uncovering the Complexities of Stakeholder Decision-Making in AI Fairness Assessment. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, 1–30.
- Nakao, Y.; Strappelli, L.; Stumpf, S.; Naseer, A.; Regoli, D.; and Gamba, G. D. 2022. Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness. arXiv:2206.00474.
- Nakao, Y.; Strappelli, L.; Stumpf, S.; Naseer, A.; Regoli, D.; and Gamba, G. D. 2023. Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness. *International Journal of Human-Computer Interaction*, 39.
- Nimmo, R.; Constantinides, M.; Zhou, K.; Quercia, D.; and Stumpf, S. 2024. User Characteristics in Explainable AI: The Rabbit Hole of Personalization? In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.
- Prolific. 2026. *Prolific*. London, UK. Version: [January, 2026], First release: 2014.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "why should i trust you?" explaining the predictions of any classifier. In *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*.
- Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K. T.; and Ghani, R. 2019. Aequitas: A Bias and Fairness Audit Toolkit. arXiv:1811.05577.
- Schuwirth, L. W. T.; van der Vleuten, C. P. M.; and Donkers, H. H. L. M. 1996. A closer look at cueing effects in multiple-choice questions. *Medical Education*, 30(1): 44–49.
- Taka, E.; Nakao, Y.; Sonoda, R.; Yokota, T.; Luo, L.; and Stumpf, S. 2024. Human-in-the-loop Fairness: Integrating Stakeholder Feedback to Incorporate Fairness Perspectives in Responsible AI. arXiv:2312.08064.
- VDAB. n.d. Sectorindeling beroepenclusters. <https://extranet.vdab.be/home/ook-voor-jou/arbeidsmarktverkenner/sectorindeling-beroepenclusters>. Accessed: 2026-07-30.
- Verheyden, A.; Zhang, Y.; De Croon, R.; Stumpf, S.; and Verbert, K. 2026. Supplementary Materials for "Spotting (and Missing) Algorithmic Bias: Investigating User Understanding in a Fairness Assessment Tool" (AIES 2026).
- Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *Proceedings - International Conference on Software Engineering*.
- Waller, M.; Rodrigues, O.; and Cocarascu, O. 2025. Beyond Consistency: Nuanced Metrics for Individual Fairness. In *ACMF AccT 2025 - Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2087–2097. Association for Computing Machinery, Inc. ISBN 9798400714825.
- Wexler, J.; Pushkarna, M.; Bolukbasi, T.; Wattenberg, M.; Viegas, F.; and Wilson, J. 2020. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26.
- Yan, J. N.; Gu, Z.; Lin, H.; and Rzeszutarski, J. M. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Conference on Human Factors in Computing Systems - Proceedings*.