# Intersectional Experiences of Unfair Treatment Caused by Automated Computational Systems

TOM VAN NUENEN*, King's College London, England
JOSE SUCH, King's College London, England
MARK COTE, King's College London, England

This paper reports on empirical work conducted to study perceptions of unfair treatment caused by automated computational systems. While the pervasiveness of algorithmic bias has been widely acknowledged, and perceptions of fairness are commonly studied in Human Computer Interaction, there is a lack of research on how unfair treatment by automated computational systems is experienced by users from disadvantaged and marginalised backgrounds. There is a need for more diversification in terms of the investigated users, domains, and tasks, and regarding the strategies that users employ to reduce harm. To unpack these issues, we ran a prescreened survey of 663 participants, oversampling those with at-risk characteristics. We collected occurrences and types of conflicts regarding unfair and discriminatory treatment and systems, as well as the actions taken towards resolving these situations. Drawing on intersectional research, we combine qualitative and quantitative approaches in order to highlight the nuances around power and privilege in the perceptions of automated computational systems. Among our participants, we discuss experiences of computational essentialism, attribute-based exclusion, and expected harm. We derive suggestions to address these perceptions of unfairness as they occur.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **User studies**; • **Social and professional topics** → **User characteristics**.

Additional Key Words and Phrases: algorithmic fairness, intersectionality, automated computational systems, conflicts

**ACM Reference Format:**
Tom van Nuenen, Jose Such, and Mark Cote. 2022. Intersectional Experiences of Unfair Treatment Caused by Automated Computational Systems. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 445 (November 2022), 30 pages. https://doi.org/10.1145/3555546



## 1 INTRODUCTION

Automated computational systems increasingly influence decisions in policing, marketing, employment, credit reporting, finance, and other opportunities. Against the belief in these systems as being "neutral", unfairness is reported across the industry [2, 10, 18, 41, 51, 109, 111]. Examples of algorithmic unfairness include facial recognition algorithms that do not recognise Black users [20], biased talent ranking systems deployed by hiring platforms [50, 64], or the mirroring of stereotypes around gender and occupation in ML-based engines [43, 70]. Systems such as these contribute to discriminatory or otherwise repressive practices, and significantly harm human rights [8, 11, 49, 108].

Authors' addresses: Tom van Nuenen, tom.van_nuenen@kcl.ac.uk, King's College London, London, England; Jose Such, King's College London, London, England, jose.such@kcl.ac.uk; Mark Cote, King's College London, London, England, mark.cote@kcl.ac.uk.


2573-0142/2022/11-ART445
https://doi.org/10.1145/3555546

Algorithmic fairness thus looms large in public policy circles, academia, and the press [11, 48]. Many responsible research approaches advocate opening processes used to train or control machine learning systems [4, 92]. How exactly to define algorithmic fairness itself, however, has been viewed differently by different stakeholders [e.g. 76, 96]. Researchers in Human-Computer Interaction have shown a keen interest in exploring these public perceptions [54, 124]. Such research tends to focus on algorithmic decision-making in particular contexts, often in order to predict fairness evaluations defined in advance [e.g. 55]. However, perceptions of fairness and discrimination are multi-dimensional and context-dependent. There is a need for more diversification in terms of investigated domains and tasks [107], and to take demographic differences into account [96].

In this work, we take an intersectional view on algorithmic injustices from the point of view of aggrieved groups and the harms they are facing from systems they interact with. We seek to understand how users who are faced with automated decisions make evaluations about their fairness and discriminatory results. To do so, we identify users from marginalised backgrounds who are especially at risk to be discriminated against. We define "at-risk" participants as those who possess vulnerable characteristics, as defined in Section 3.2.1; "intersectional" participants as those who possesses combinations of at-risk categories; and "privileged" participants as those who self-identify as white, straight, cisgender, non-disabled and from the Global North. We use the Critical Incident Technique to ask these participants about their own experiences with unfair decision-making, and take an intersectional approach to analyse the compounded social inequalities we find, engaging in qualitative analysis of responses and a quantitative analysis of prevalence (i.e. proportion) of unfair and discriminatory experiences among these groups. Our main research questions are:

(1) What is the prevalence of perceived unfair treatment and discrimination between privileged and intersectional participants?
(2) Are there significant differences between privileged, at-risk, and intersectional participants with regards to the reporting of unfair treatment and discrimination?
(3) How do privileged and intersectional participants discuss different types of unfair experiences?
(4) How do privileged and intersectional participants discuss different types of unfair automated computational systems?
(5) How do privileged and intersectional participants describe discriminatory harm?
(6) What kinds of strategies do privileged and intersectional participants employ to reduce harm?

## 2 RELATED WORK

Algorithmic fairness is a contextually situated concept that touches on debates in political philosophy around equal treatment and distribution [15]. Discrimination, a related but more legally specific term, assumes the unequal treatment of a person based on certain protected features such as ethnicity, disability status, or gender [15, 32, 91, 121]. *Disparate treatment*, in this context, is the intentional or direct application of procedures to one individual compared to another based on protected attributes. Algorithmic decision-making, in contrast, is more typically characterised by *disparate impact*, in which "an apparently neutral provision, criterion or practice would put persons of a protected ground at a particular disadvantage compared with other persons" [126, p.4]. Disparate impact can take distributive and representational forms [15, p.8], where the former refers to unequal resource distribution, such as in the automated allocation of loans or the setting of insurance premiums. The latter includes the fair representation of identities, cultures, ethnicities, languages, or other social categories.

### 2.1 Fairness in computer science

As Verma and Rubin show in their overview of fairness definitions [117], technical research on algorithmic fairness often takes on a prescriptive form, focusing on the evaluation and optimization of formal definitions of fairness, such as positive predicted value or false positive rates. Treating fairness as a mathematical optimisation problem has well-known limitations, and there is significant incompatibility between the fairness conceptions offered in different studies [73]. Narayanan et al. [89] detailed up to 21 existing fairness models, stressing that definitions are weighted with values and politics, and that mathematical definitions imply substantial moral assumptions. Different fairness trade-offs have been proposed, with multiple authors highlighting the importance of social context when assessing appropriate understandings of algorithmic fairness [42, 78, 123].

There is a broadly shared understanding that algorithmic decision-making systems need to be legitimised from public perspectives [55, 80, 98]. Survey-based work often tries to evaluate or predict fairness evaluations by different stakeholders, and is connected to work in Explainable AI [e.g. 61, 105]. Research of this kind tends to focus on particular application domains and scenarios, such as pretrial risk assessment and hiring decisions [107]. For instance, Grgic-Hlaca et al. have explored how users perceive and reason about fairness in algorithmic decision making using scenario-based surveys [55]. One limitation of these empirical studies is that they either use simple single-item measures of perceived fairness, or adopt fairness scales that were initially designed for human decision-making [e.g. 29]. Yet, preferences for different distribution norms seem highly context-dependent and can vary substantially across domains, tasks, or demographics [96]. People who experience marginalisation perceive human and computational decision-making differently; as such, there is a need to purposely recruit and study different social groups and different dimensions that can account for individual differences in experiences with AI [77].

This demonstrates the need for more diversification in terms of the investigated domains and tasks related to algorithmic fairness [107].

### 2.2 Critical approaches and Intersectionality

The importance of context for perceptions of fairness is central to the fields of science and technology studies (STS) and critical algorithm studies, which articulate the heterogeneous influence of algorithms in society [e.g. 10, 38, 84, 86, 124]. This work, as well as related studies in moral and political philosophy, points out that fairness always implies a political choice [123]. Algorithmic fairness approaches cannot apply a single egalitarian calculus across different social contexts, but should also consider how inequality is produced [15]. Instead of attempting to "unbias" technology, these approaches account for the asymmetrical and historically embedded power relationships that support its development and deployment [13, 40].

Data feminism, critical race studies, and data activism offer related lenses to approach computational systems in terms of structural or systemic racism, heterosexism, ableism, and other compounded prejudices. These approaches show that "big data" not only refers to anonymous, decontextualised large datasets used by institutional and corporate actors, highlighting instead that data always operates as part of knowing, perceiving and sensing human bodies — in other words, "how data comes to matter" [79]. In doing so, these approaches oppose the trend of data universalism, that is, portraying technology and datafication as unrelated to history and unique sociopolitical, cultural, and economic settings. They also complicate research design approaches in HCI where participants are tasked with evaluating imaginary scenarios they do not know first-hand, e.g., by comparing models for deciding whether to grant bail to criminal defendants [e.g. 55, 60].

"Studying up" on these critical approaches [7] means that injustices need to be viewed from the point of view of the aggrieved groups and the particular harms these groups are facing, and

taken as a non-exclusive but valuable and essential source of knowledge [57]. A central concept here is *intersectionality*, the intersection of multiple emerging and compounding inequalities that make people more susceptible to emotional, financial, physical harm, neglect, or discrimination in society [82, 120]. The concept, originating in the black feminist movement and popularised by Kate Crenshaw in the 1980s [28, 30], has recently began to be adopted in human-computer interaction research and computer science [e.g. 20, 82, 112]. It is a powerfully evolving analytical framework to recognise (systems of) oppression, acting as a powerful concept in critical social theory [28, 85].

The intersection of "at-risk" characteristics poses a significant challenge for AI and ML in terms of both their society-wide application and scholarly analysis. The understanding of identity characteristics as multiplicative factors of inequality needs to be considered in conjunction with the increasingly granular data used for decision-making by predictive systems, creating a range of vulnerabilities. In the next section, we detail our methodological approach when accounting for and listening to these marginalised individuals.

## 3 METHODOLOGY

We enabled research participants to define and explain unfair treatment in their own words, instead of using definitions and explanations deriving from previous research, which are then incorporated into a survey instrument. We thus aim to minimise the definitions and explanations of fairness in AI derived from previous research into our survey instrument [57]. To gather experiences of unfair treatment from users we borrow from the Critical Incident Technique (CIT) [21, 44], as detailed below. As the production of knowledge is itself shaped by a researcher's positionality [68], the methodological choices we describe here were extensively discussed with colleagues from different backgrounds. This includes the technical domains of AI and Human-Computer Interaction, critical social studies on Data Feminism, and Critical Algorithm Studies.

### 3.1 Survey Instrument, Ethics, and Data Quality

The CIT seeks to find factors that help and hinder participants in relation to a specific phenomenon — in this case, dealing with forms of algorithmic harm as they are experienced in everyday life. The technique has been used across disciplines, including communications, nursing, education and teaching, medicine, marketing, and psychology [21]. The guidelines for the CIT were followed to ask for the most recent observation (i.e., the last time participants experienced unfair treatment). This is in order to prevent biasing the study to just the more dramatic or vivid incidents [118]. We believe this approach to be valuable as it provides descriptive (instead of prescriptive) insight into people's holistic experiences with the output and integrated nature of algorithmic systems. Our work is inspired by previous studies which have successfully used a similar methodology to collect and study specific experiences in the use of software updates [116], as well as reported experiences about security [97] and privacy [110].

In order to administer the survey we used Prolific, an online participant recruitment platform, paying participants according to minimum UK wage. We acknowledge that having vulnerable individuals share their experiences entails a risk of epistemic exploitation, which occurs when privileged persons compel marginalised persons to educate them about the nature of their oppression [14, 45, 95]. We should be careful to ensure that ethical research design contributes towards social justice [106]. Instead of demanding victims to do the emotionally exhausting work of reliving their experience and defending their interpretation of it, we ask open questions without necessitating answers, and allow participants to stop at any time without this preventing being paid.

The BDM Research Ethics Panel at our institution approved this study. Participants were not identified through their Prolific ID and we only saved pseudonymised data. Participants could withdraw from a study by clicking "Cancel Reservation" before the study began, or "Stop without

Completing" on Prolific during the study. This did not affect their Prolific score or payment in any way. Partially completed data was, obviously, not used.

After having participants sign a consent form, we first asked a set of closed- and open-ended questions about the last time participants experienced unfair treatment from automated computational systems based on personal features — see the full questionnaire in Appendix A. Before running our survey, we ran a pretest with 50 participants, resulting in a revision of ambiguous or unclear questions. Our central question was:

> "When was the last time you felt to be treated unfairly when interacting with an automated computational system? Recall: we are talking about an automated system (not another person) treating you unfairly because of your personal details."

We chose to prime our participants as little as possible: however, at the beginning of the survey, we did mention online service, computer, application, or smart devices as examples of automated systems. Similarly, we mention "gender, age, sex, religion, disability, etc. or social groups or communities you consider yourself to be a part of" as examples of personal details. We also did not distinguish between voluntary and compelled usage of systems, in order to let participants decide what kind of experience they themselves found most memorable. Our questions were semi-structured following the known stages of conflicts: conflict identification, communication, and resolution [37]. In terms of resolution, we asked whether participants altered their usage of the computational system in question (e.g., contacting the owner of the system or platform, deleting their profile, changing their privacy settings, and so on) and why. We also asked whether participants who had experienced unfair treatment whether they considered it discriminatory, and why.

In the second part of the questionnaire, we asked participants about their perceived ability to engage with computers, IT services, and the Internet. Previous research has shown conflicting results when connecting these technical literacies to people's perceived notions of fairness [3, 76]. Finally, we asked a series /textcolorredof demographic questions around participants' protected characteristics including gender, sexual preference, ethnicity, disability status, and location.

With regards to data quality measures, we added one attention check in the first (and most qualitatively heavy) section [93]. In order to minimise straightlining [72], which we recognise itself can be a response to the emotionally taxing work of recounting discriminatory harm, we avoided matrix questions as much as possible and benefited from different (as opposed to homogeneous) types of closed questions. We checked manually for straightlining in our initial data curation by applying simple non-differentiation methods for the closed questions [72], as well as checking for repeated or overly brief answers in the qualitative sections (e.g. "blabla").

### 3.2 Recruitment Approach, Intersectional Tenets, and Demographic Analysis

The average Prolific user skews white, male, US and UK-based, and cisgender. However, Prolific allows researchers to select participants using 100+ demographic screeners. Not only does this help to reduce selection bias, it allows us to give weight to the intersectional aspects of the unfair treatment reported by our participants by counting different demographic features that, based on previous literature, we consider to increase discriminatory risk. We proceeded in batches, oversampling based on a subset of values for certain protected characteristics, namely gender (non-binary), sexual orientation (non-cisgender), ethnicity (non-white), disability status, and location (developing countries). We thus ran five different questionnaire rounds with 130 participants each.

*3.2.1 At-risk categorisation.* Using dummy variables, we defined an "at-risk" participant as someone who possesses one of the features discussed below. Our motives for classifying these at-risk characteristics are as follows:

- **Gender**: We counted all participants who identified themselves as being "beyond the binary" of male/female towards our at-risk category (e.g. non-binary, fluid, transgender, genderqueer, and gender nonconforming). Given well-documented discrimination against transgender people and the structural inequities which contribute to these people's invisibility [71, 100], we counted participants identifying as transgender as another separate at-risk category.
- **Sexual orientation**: we considered self-identifications beyond "straight" to constitute an at-risk characteristic. This includes participants who identified as lesbian, gay, bisexual, transgender, queer, or another non-privileged category. Such LGBTQ+ adults in the United States experience pervasive discrimination across many areas of life, including widespread interpersonal manifestations [23].
- **Ethnicity**: we considered participants from non-white ethnoracial backgrounds to be at-risk. Our participants all come from countries in which non-white backgrounds are discriminated against. Of note here is the fact that our African respondents are all from South Africa, where Black groups have been shown to perceive chronic discrimination as well [122].
- **Location**: We compared countries that appear in the "Developing economies" list of the UN's World Economic Situation and Prospects 2020 [114, p.166], which is loosely equivalent to what is considered the "Global South". We categorise participants from this list as at-risk, as vulnerable Southern populations are at risk of discriminatory harm from AI that is designed with little regard to local context [25, p.604]. We are aware that the Global South is a composite and plural entity within which privilege is distributed unequally [66, 84], and our approach precludes stratified privilege within national context — e.g., white South Africans tend to be more privileged than Black South Africans.
- **Disability**: We consider disability status an at-risk variable. AI makes significant interventions in disabled people's lives, yet relatively little work has been done to focus on disability and AI [13, 113]. As disabled persons have been historically disadvantaged and are more likely to be underprivileged, there is a need for inclusive, participatory and value-sensitive design to center their marginalised perspectives [113].

Table 1 provides a breakdown of participants with aforementioned at-risk characteristics. Because of our recruitment approach and categorisation, the majority our sample consists of participants that we classify as possessing at least one at-risk characteristic.

*3.2.2 Intersectional categorisation.* We use combinations of at-risk categories to determine intersectionality. Using this method, we classified 29% (N=180) of our participants as intersectional. Further, in order to compare harms between marginalised and privileged groups, we classified 14% (N=88) privileged participants as those who are white, straight, cisgender, non-disabled and from the Global North. The remaining 56% (N=343) we classify as at-risk.

Instead of "proving" intersectional harm through statistical relevance scores alone, we are interested in what characterises them, and to oppose the experience of intersectional participants to those of the *privileged* group we establish — that is, participants who are white, heterosexual, cisgender, non-disabled, and from the Global North. We also aim to situate the experiences of both groups in the context of systems they mention as being unfair.

Our analysis is largely qualitative, and aims to pay attention to the fact that oppression is a system that works to maintain privileges, and occurs simultaneously at individual, institutional, and cultural levels [33, 58]. While we are interested in analysing the particularities of our intersectional group, we recognise that positionality does not determine disadvantage in any simple, additive way [9, 22, 56]. Intersectionality should first of all be considered as an "analytic sensibility" [26]—that is, a theoretical framework that requires us to avoid assuming homogeneity across intersections both in outcomes and processes, and to challenge the notion that independent variables can be

| | | N | % |
|---|---|---|---|
| Ethnicity | American Indian or Alaska Native | 5 | 2% |
| | Asian | 59 | 19% |
| | Black or African American | 72 | 23% |
| | Native Hawaiian or Pacific Islander | 3 | 1% |
| | Other | 178 | 56% |
| Sexual orientation | Asexual | 4 | 3% |
| | Bisexual | 84 | 53% |
| | Demisexual | 2 | 1% |
| | Gay | 29 | 18% |
| | Lesbian | 26 | 16% |
| | pansexual | 9 | 6% |
| | Queer | 10 | 6% |
| | Questioning or unsure | 6 | 4% |
| | Asexual | 1 | 2% |
| Gender | Fluid | 5 | 8% |
| | Gender neutral | 4 | 6% |
| | Non-binary | 48 | 76% |
| | Questioning or unsure | 3 | 5% |
| Transgender | Yes | 36 | 57% |
| | Questioning or unsure | 27 | 43% |
| Disability | Yes | 44 | 75% |
| | Questioning or unsure | 15 | 25% |
| Location | Africa | 57 | 28% |
| | Asia | 2 | 1% |
| | Latin America | 137 | 68% |

Table 1. Descriptive statistics of at-risk participants

neatly separated [85]. However, we believe the provisional use of categories as anchor points can still be valuable, even though these points are not static [53, p.14]. We considered intersections of being racialised, sexual minoritised, gender minoritised, disabled, or living in the Global South.

For the purpose of this study, we decided to forego predictive statistical tests such as logistic regression with interaction terms. This is partly due to the small sample size we are working with for particular (interactions of) features, as well as issues of multicollinearity. While the latter can be tackled by dropping or combining features, we believe that building our current analysis on the results of such regression models would imply treating intersectionality as a "testable explanation". Approaches of this kind generally employ a standard positivist empirical examination to test the claims regarding discrimination or lack of access asserted by normative intersectionality theorists [56, p.268]—for instance, by including additional variables and interaction terms to determine whether race, gender, and other factors play a role in predicting some outcome. While instrumentally valuable, these approaches venture quite far from the theoretical tenets of intersectionality that we aim to highlight in this paper—particularly, the qualitative difference in responses between intersectional and privileged respondents. Given this focus, we limited ourselves to a series of

chi-square tests of independence between our at-risk features, to see whether correlations could be found that might be explored in further studies.

### 3.3 Analysis

We analysed answers to the closed-ended questions in our survey instrument — for instance, whether a system's treatment was considered discriminatory, or to what extent the experience was found to be hurtful — by performing standard statistic analyses, particularly chi-squared tests for independence for categorical responses and correlation coefficients for ordinal (Spearman) and ordinal vs categorical (point-biseral) responses. We also corrected for multiple tests using Benjamini-Hochberg with a 10% false discovery rate [12], which is known to strike a good trade-off between Type I and Type II errors [39].

Our approach to the open-ended questions was through coding responses using an inductive, open coding approach. Here, we followed methods from grounded theory, which involve systematic, yet flexible guidelines for collecting and analysing qualitative data to construct theories "grounded" in the data themselves [24]. Grounded theory uses a data analysis procedure called theoretical coding to develop hypotheses based on what research participants say [5, 101]. We argue that, especially in the case of potential intersectional harms, this iterative and reflexive method would allow the researchers to reflect on participants' narratives and concerns. We believe there is a need for such grounded approaches in computer science, in order to engage in what Davis calls "empathy work" [35].

Coding of the qualitative data took place in two rounds. Following grounded theory [52], the main researcher first began the coding process after the survey was completed. Coding was started early to identify interesting codes and categories that could be explored, leading to descriptive classes of both the type of unfair treatment and the type of systems that respondents reported on. We created a codebook consisting of the categories that summarised the data most usefully [44, p.344]. We then tested the reliability of these codings with a sample of 50 respondents that were assessed by another rater from a different ethnic and gender background. Following Higgins and Deeks's recommendation, we calculated the consistency measure Cohen's Kappa ($\kappa$) to ensure coherence of the coding results between the two raters. The first coding process resulted in a Cohen's $\kappa$ value of .64, which reflects "good agreement" [63, p.155]. To improve the reliability, both raters discussed the examples on which they diverged and refined the codebook. In the next step, a second test with 50 abstracts was conducted, leading to a Cohen's $\kappa$ score of .77, which is considered "excellent agreement".

These codes were then aggregated into broader, thematic codings, which were more directed, selective, and conceptual. In this phase, we used the same codes for answers to all of our qualitative questions, as we found that participants took note of important themes at different points of the survey. We used a maximum of three most-relevant categories per answer, as we found this captured nearly all of our data. We discuss these broader themes in the discussion, relating them to broader questions to be asked about unfair and discriminatory treatment.

## 4 RESULTS

### 4.1 Prevalence

In order to answer RQ1, we note that 663 participants submitted the questionnaire. From those, we removed 9 participants who failed at least one ACQ, 24 participants who did not finish the survey, and 24 participants who were straightlining or answering with nonsensical text in the open-ended questions. From the valid 611 participants, 320 (that is 52%) reported having experienced unfair treatment from automated computational systems. From the 320 participants who reported

| | | Unfair treatment | Discriminatory treatment |
|---|---|---|---|
| Ethnicity at-risk | N | 609 | 294 |
| | p | .378 | **.001** |
| | $\chi^2$ | .778 | 11.196 |
| Sexual orientation at-risk | N | 611 | 294 |
| | p | **.003** | **<.001** |
| | $\chi^2$ | 8.916 | 18.052 |
| Gender at-risk | N | 610 | 294 |
| | p | .563 | **.021** |
| | $\chi^2$ | .334 | 5.363 |
| Transgender at-risk | N | 609 | 294 |
| | p | **.008** | **.003** |
| | $\chi^2$ | 6.954 | 9.119 |
| Disability at-risk | N | 609 | 294 |
| | p | **.014** | **.024** |
| | $\chi^2$ | 6.094 | 5.091 |
| Location at-risk | N | 603 | 291 |
| | p | **.000** | **.031** |
| | $\chi^2$ | 21.744 | 4.661 |
| Intersectional | N | 611 | 294 |
| | p | **.005** | **.002** |
| | $\chi^2$ | 7.811 | 9.277 |
| Privileged | N | 611 | 294 |
| | p | .210 | **.000** |
| | $\chi^2$ | 1.569 | 17.700 |

Table 2. Chi-squared scores for at-risk groups vs all other groups, intersectional vs all other groups, and privileged vs non-privileged groups, as related to perceived unfair and discriminatory harm. Bold means *significant* after correction for multiple tests (cf. Sect 3.3).

having experienced unfair treatment, 163 participants (51%) considered their experience to be discriminatory, 132 participants (41%) did not consider their experience to be discriminatory, and 26 participants (8%) were unsure (although more participants indicated some form of uncertainty around their answer — see Section 4.4). We did not find a significant correlation between privilege and the reporting of unfair treatment ($\chi(1, N = 611) = 1.569$ and p=.210); however, this reporting was correlated with intersectional characteristics ($\chi(1, N = 611) = 7.811$ and p=.005).

In order to answer RQ2, we first traced correlations in our data by performing a series of chi-squared tests of independence to examine the association between perceived unfair treatment and the at-risk groups we identified in Section 3.2. As Table 2 demonstrates, the at-risk groupings resulted in more significant[1] correlations than chi-squared tests performed over the original variance in protected characteristics. Results indicate significant correlations between most at-risk characteristics (vis-a-vis not possessing that characteristic) and the experience of unfair and discriminatory treatment. We also looked at correlations between the experience of unfair and discriminatory treatment and participants' having an intersectional or privileged background (vis-a-vis not having an intersectional or privileged background). Results indicate significant correlations between an intersectional background and the experience of both unfair and discriminatory treatment.

[1]Recall that, as explained in Section 3.3, we used Benjamini-Hochberg to correct for multiple tests.

We were also interested in participants' self-reported IT literacy as an additional feature that might explain perceptions of unfair treatment. We thus asked respondents to score themselves with regards to their technical literacy skills. We calculated the Cronbach's Alpha to measure the internal consistency of the different literacy questions we asked: the value was .733, which corresponds to "acceptable". Results show that these skills for our participants skew positive. However, participants' overall literacy value was not found to correlate with the reporting of unfair or discriminatory treatment.

### 4.2 Types of unfair treatment

In order to answer RQ3, we produced separate bottom-up codings of unfair experiences yielding over 20 types of unfair treatment mentioned by users, and over 20 types of unfair systems. We focus here on the top-10 categories for both variables. Figure 1 summarises the 10 most-common types of unfair treatment mentioned by users for our privileged, at-risk and intersectional groups. We will first discuss these experiences, after which we will zoom in on the system contexts in Section 4.3.

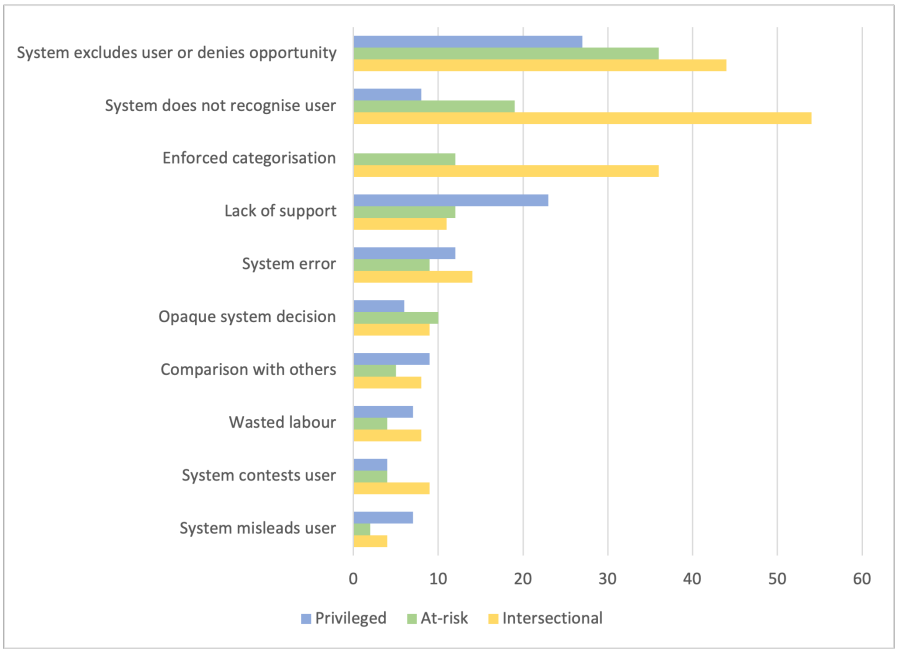


Fig. 1. Experiences of unfairness for different participant groups

*4.2.1 Exclusion and recognition.* The most frequently mentioned problem overall is involves being excluded from some access, service, or broader opportunity by a system ("system excludes user or denies opportunity"). In these cases, unfairness is often connected to an unfair algorithmic weighing of variables. For instance, when discussing being rejected by a government hiring system, one participant noted:

> P84: "I noted only my grades were recognized by the algorithm. [It was unfair] because it mainly evaluated my grades from my career and obviated my practical experience, which in my opinion it should be more important."

While these issues are discussed by all groups, they become especially problematic when they are caused by protected attributes. One participant with an intersectional background wrote:

> P203: "I have noticed that answering 'Other' for gender identity is significantly more likely to cause me to be "screened out" of a survey or gig work opportunity, simply because the algorithm is looking, for example, for a certain number of 'men' and 'women', and so chooses to discard my input automatically—even though the content of the research or gig opportunity has nothing to do with, and is not at all affected by one's gender."

The response demonstrates the precariousness of gig economy labour, in which work is reconfigured for optimal human capital extraction [27, p.32]. Workers in these systems are repeatedly selected and filtered based on their protected characteristics — including in our own study. Of course, these exclusionary practices can be partly explained due to the legitimate need of researchers to target particular demographics. What needs to be highlighted is that many respondents view these exclusions as a matter of disparate impact, as they are based on features that in typical employment settings cannot factor into a hiring decision due to equality legislation.

The division of labour through discrimination is key to the production culture of microwork systems in particular [67], and is connected to geopolitical dimensions of microwork, which often entail citizens from the Global South doing manual labour for employers in the Global North. This is underscored by the fact that microwork platforms are frequently mentioned by participants from the Global South, and a considerable amount of users felt excluded by microwork platforms based on their location (N=23). For instance, one participant discusses working for an SEO service paying them to click websites, noting that:

> P223: I get paid the half of the amount a U.S. based person would get just because I'm mexican, even though we do the same job.

With many of these platforms cutting across borders and covering a global marketplace, the scope of what people expect as "fair" decision-making, and their ideas about equal opportunity, play out on a global level as well. This highlights the intractability of the "global crowd ethics" that accompany microwork platforms [67, p.734].

At the same time, P203 points at a concern about features not being relevant to the system at hand, which Grgic-Hlaca et al. describe as a latent evaluation property of "feature relevance" [55]. Several of our participants highlight that microwork is not simply stratified by location, but that the inclusion of irrelevant features causes further exclusion among minorities within those locations. This is just one example of how intersectionality can identify compounded harms in algorithmic contexts.

A related, and commonly experienced issue with automated systems is one of visibility, where the user is not excluded by being recognised as being from a particular class, but by not being recognised at all ("system does not recognise user"). We see that intersectional participants are over-represented in this category (see Figure 1). One participant wrote:

> P144: "i applied and started a study on disability at work. during the screening questionnaire, a question came up asking me to name the type of disability i have. it did not include options for depression, anxiety disorder or panic disorder. i could not move forward unless i said that i did not have a disability, which was not true."

Beyond their disability not being recognised, the participant shows that the recognition and valorisation of an identity label that may encourage people to force themselves to fit into a category to which they do not belong [34, p.407]. This leads to the next category.

*4.2.2 Categorisation and contestation.* The experience by P44 is also an example of the experience with the highest relative frequency among intersectional participants, namely being forced into an inaccurate or adjacent class or category by the system ("enforced categorisation"). The theme indicates a common lack of specificity in algorithmic systems: they do not have enough options, or cannot compute the specificity of the user. This issue of categorisation is also related to predictive recommendations and what we could call algorithmic funnelling. Participants commonly are unduly pushed into categories by such recommendation systems.

> P299: "'I follow quite a few LGBTQ+ creators on youtube, so I would expect that my home and explore pages on youtube would have a number videos from said creators or have videos with the same type of content. However, for whatever reason, when I look at my home and explore pages, there's much less LGBTQ+ content being shown to me than I would expect. Even in my subscriptions page, which should be filled with the videos from the creators I subscribe to, sometimes the LGBTQ+-focused videos are missing, even when other videos from the same LGBTQ+ creators are being shown."'

The participant's analysis points to the pervasiveness of algorithmic folk theories:"intuitive, informal theories that individuals develop to explain the outcomes, effects, or consequences of technological systems, which guide reactions to and behavior towards said systems" [36, p.3165]. Especially in the context of social media, these perceptions are based on algorithms operating on the basis of identity, the worry being that these systems end up valuing some identities over others [69]. Interestingly, this not only means that particular identities are suppressed, as is noted by the following participant:

> P253: "Watched a video about a mental disability,now all my reccomended [sic] videos are about that.Its stressing since i suffer of bypolar [sic] disorder"

Indeed, the issue of being algorithmically targeted not only pertains to a lack of representational options, but an oppressive granularity combined with prediction. As one participant noted:

> P226: "On Okcupid there's an option to select if I'm monogamous or polyamorous and even if I'm trying to look for people to chat with I can't get all the results (in other words, I get only monogamous if I choose monogamous and vice versa). Also, if I want to gender myself as non-binary (on both mentioned sites) I will only get matched to people looking specifically for non-binary (which defeats the purpose of being agender). This causes me to lie in order to even have a chance on getting matched with the gender I prefer."

The participant points to feature volitionality [55], insofar as they feel trapped in the stereotypical perception of their sexual orientation, which is non-volitional. Self-categorisation here is linked to predictive recommendation, and this prediction is simplistically mirroring input characteristics. This is a significant issue for intersectional users such as P226, as their identity is automatically being taken for their preference. The recognition systems under discussion here implicitly assume that sexual preference is a static concept that does not frequently change across time and cultures [48]. Again, we have to note here that oppression for some groups is interconnected with opportunity for others: this predictive mirroring may be a convenience when it pertains to a privileged identity.

We created a separate category for the cases in which systems were found to be actively contesting users, or changing their input ("system contests user"). One participants discusses an automated tagging system that contests concomitant human tagging:

> P292: "[My job] involves tagging content according to subjective characteristics, and then write and explain your reasoning for such tags. Then, you get scored by an automatic system that does not care about what you wrote nor even reads the comments,

> and simply marks some of the opinions you gave - which they asked you to provide and told you it was part of the job - as 'wrong'."

The fairness judgment made in this comment is related to the system's perceived inability to reliably assess an outcome [55]. In the context of human work, the participant's fairness evaluation relates to a sense of purposelessness, as their work is being evaluated by an automated Machine Learning system, which is nonetheless (and perhaps ironically) trained on the same human tagging. This feeling of labour being made undone by a system ("Wasted labour") returns throughout the data. The comment is also indicative of a theme of being misled ("System misleads user") by a system and its designers — in this case, as the employee believed that their work would end up informing the system's decisions, instead of the other way around.

*4.2.3 System opacity and interpersonal comparison.* The lack of specific knowledge about these systems is evidenced in our data by the amount of people indicating being unsure how the system made its decision ("Opaque system decision"). This opacity is a feature of AI systems in particular, and can lead to the "algorithmic folk theories" mentioned above. One respondent who denied a job by a recruitment system wrote:

> P14: "I don't have the exact proof, it is not obvious. But I think that (and have read about) there are recruitment systems which filter out candidates with gender or ethnic background bias. "

It is arguably due to this system opacity that many participants tentatively compare their treatment with that of others ("Comparison with others"), pointing to an awareness of distributive fairness [15]. For instance, one of our intersectional participants noted:

> P48: "i applied for financial assistance but was declined, a trusted friend with similar profile but diffarent [sic] race and gender were approved"

Again, while there is no way of proving whether the anonymous system mentioned here was discriminating based on ethnicity, the response indicates that people in historically disadvantaged positions are oriented towards fairness and its relation to discrimination. These remarks align with latent fairness properties regarding outcome disparity and sensitive group membership [55]. Our data also demonstrates that people often come up with speculative answers with regards to fairness in automated and AI systems — a finding that has been reported on in previous consumer polls [94]. When trying to get answers to their questions, we see participants being frustrated when this is not possible ("lack of support"); we discuss this in Section 4.5.

### 4.3 Types of unfair systems

In order to answer RQ4, we coded the different kinds of systems that users mention as being unfair, in order to see how participants' experiences are situated in particular contexts. These are summarised in Figure 2. Several systems and experiences immediately jump out: particularly, a sense of exclusion in microwork platforms, social media and recruitment platforms, a lack of support in support systems, occasions of enforced categorisation in registration systems, and a wider sense of not being recognised as across systems.

*4.3.1 Microwork platforms and recruitment platforms.* The most commonly experienced unfair system were reported around microwork platforms such as Prolific or Fiverr, as discussed in Section 4.2.1. Given we used Prolific for recruitment, these findings could be expected. The narratives our participants share involve being screened out of opportunities or studies based on their gender, ethnicity, or location, showing that oppression for some groups is interconnected with opportunity for others [85]. See also P203 and P223 in Section 4.2.1.

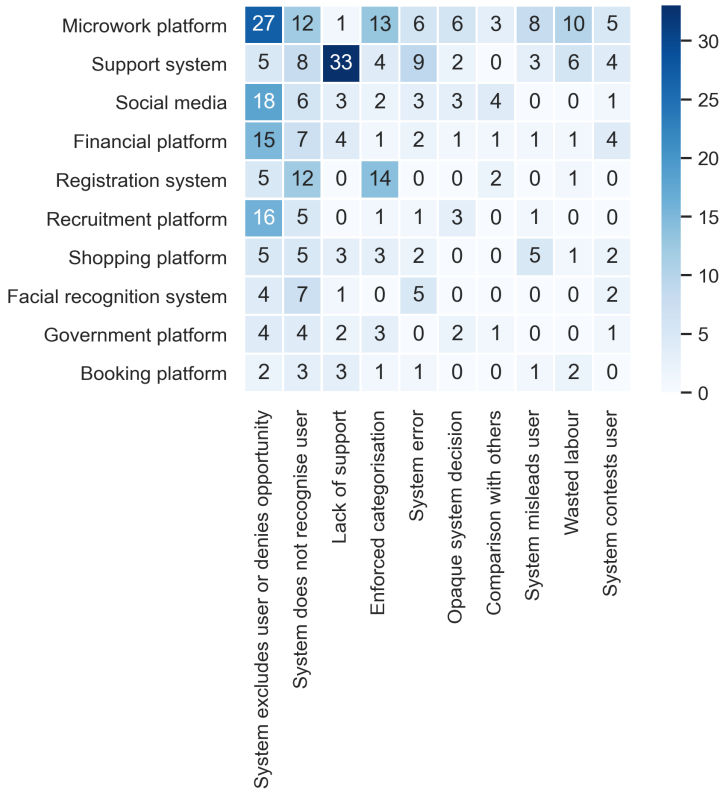


Fig. 2. Heatmap of top 10 experiences of unfairness in top 10 systems/platforms

*4.3.2 Support systems, Social media and Video platforms.* Support systems, which provide help or assistance to users — such as chat bots and interactive voice response systems — are notably more frequent in the narratives of our privileged participants. Participants most frequently take issue with these systems due to their ambiguity and minimal feedback, which have been diagnosed as issues before [72]. More problematically, participants also report how these systems mischaracterise them by not taking important attributes into account. For instance, one participant notes:

> P42: "I was looking to get an appointment to get a clinical study for a fertility treatment, I'm a lesbian so my spouse is a woman, but the chatbot was treating us like a heterosexual couple."

Social media and video platforms such as Facebook, YouTube, and Netflix are similarly mentioned due to the algorithmic recommendation systems they incorporate (see Section 4.2.1).

*4.3.3 Financial platforms, Shopping platforms, Government platforms and other Registration systems.* Registration systems include those systems that allow users to set up an account. Firstly, financial platforms are those dealing with loans, trading, and other financial services. Participant discussions of these platforms bring us back to the issue of exclusion: 71% of the participants mentioning these systems were from the Global South (the highest percentage for all systems). Within this demographic, the distributive harms of these platforms often pertain to a lack of inclusivity. One of our participants located in South Africa noted:

> P130: "The automated system declined my application essentially because I am transgender. I was refused the layaway service because my details include a change of name, and the system was incapable of rectifying this."

The unfairness reported on these financial support systems highlights the ways in which algorithmic systems constrict disadvantaged people's opportunities even further because of their intersectionality. It shows how shopping platforms, which the above layaway example is an example of as well, cause differential harms across privilege lines. Narratives of privileged participants on these systems mainly involved inconveniences such as paying for a product that the system inaccurately depicted as available.

Participants often comment on the opacity on part of these registration systems, and suggest biases on part of system designers in terms of how attributes such as gender are made actionable for users. One intersectional participant discusses a government registration system:

> P8: "I had to change some personal information about a government procedure and they asked me to fill in information. By filling them in, it only gave me options to identify myself in ways that I did not identify myself in and did not feel represented. Because i felt i wasn't being represented from who should represent me and tke [sic] care of me [...] They didn't take into account other options in gender than man or woman and also because of that in mails they addresed [sic] me Mr."

The participant notes how the "symbolic annihilation" by their government [69] was particularly painful given the expectation of fairness in these systems. The response also demonstrates that exclusionary design choices around personal attributes can lead to longer-term mischaracterisation and classification, and thus repeated experiences of harm.

*4.3.4 Facial recognition systems.* Our last category of facial recognition systems are related to well-known forms of algorithmic discrimination: substantial disparities have been reported with regards to gender and ethnicity in the accuracy of facial recognition systems [20, 46]. Our participants mainly report unfair treatment towards ethnicity and gender in these systems, noting systems that either falsely recognise a user, or do not recognise a user at all. One participant gives an example of not being recognised by a system, suspecting racial exclusion and causing significant harm.

> P197: "I went for a passport renewal online and my picture [because of my skin complexion] failed as they was not 'enough light' in the photo despite there being plenty I had to shine light on my face to get it to work."

### 4.4 Discriminatory treatment

In order to answer RQ5, we asked respondents to explain whether they thought the experience they had could be classified as discriminatory. 41% of our participants (N=131) did not think their experience discriminatory, while 51% (N=163) did. Further, intersectional participants report discriminatory treatment significantly more often than privileged ones: 67 out of 92 (73%) for the intersectional group, as compared to 40 out of 103 (39%) for the privileged group. Using a chi-squared test of independence, we found an association between intersectionality and the reporting of discriminatory treatment, with $\chi(1, N = 294) = 9.277$ and p=.002. One intersectional participant points out how representational exclusions of minoritised people matters in the use of registration systems.

> P69: "Yes of course it was [discriminatory]. Non-binary and genderqueer people exist! We do not fit into these limited binary options and excluding us is discriminatory because people whose gender is binary do not experience that exclusion."

We also note that 8% (N=26) of our participants were unsure about whether the treatment was discriminatory or not. This is largely due to the opacity of the discriminatory systems, which cause distrust among our participants. One respondent theorises:

> P102: "To my knowledge, many companies use automated software in their hiring process. [...] I have been rejected after hundreds of applications and I strongly suspect it is because of the combination of a very competitive job market (due to COVID) as well as this automated software."

This expectancy of discriminatory harm is worrying, as it could contribute to a blanket distrust of automated and AI systems [31]. There is a clear need for more transparent communication around the occasions when automated systems are used, and how stringently their results are used.

Another salient theme we note here pertains to 11% (N=35) of our participants, who noted they did not view their treatment as discriminatory precisely *because* the system was automated. For instance, one intersectional participant noted:

> P34: "No, I do not think it was discriminatory as it was not intention. However, I do believe that it was a miscalculation on the developer/operator's part."

This misunderstanding of disparate treatment versus disparate impact highlights the need to communicate with users around machine learning biases as the byproduct of structural organisational practices and cultural norms [7].

### 4.5 Strategies for harm reduction

To answer RQ6, we first asked participants whether they took action after the incident. 133 (42%) participants noted to have acted after the harm was done, while 187 (58%) did not. However, we did not find any significant correlations between participants possessing at-risk or intersectional characteristics and taking action. We also asked whether participants' thoughts on the system that treated them unfairly had changed. Most participants (N=173, 54%) indicated having changed their view, while 109 (34%) participants did not change their view on unfair systems, and 37 (12%) participants were unsure whether they did. Again, we did not find any significant correlations between a change of view and intersectional status.

A frequent theme we identified relates to participants who note that they are disappointed in the designer instead of the automated system. One participant wrote:

> P127: "It is never my view on automated system but the owners for not using a larger sample size".

This participant understanding of training bias underscores the need for designers consult with affected stakeholders and the general public during the initial phases of system design [40]. This may especially be relevant for individuals who are disappointed in automation as a whole:

> P33: "It's made me thing that it's not such a great technological advancement after all, if it's only going to benefit the same people who already have and have had all the benefits so far."

This comment relates to another noteworthy category of expected harm, which is largely expressed by intersectional participants. This was the only response for which we found a significant relation with intersectional status through a chi-squared test ($\chi(1, N = 320) = 6.069$, p= .013). These participants explain being used to unfair treatment, and whose fatigue about these systems prevents them from seeking help. One intersectional respondent notes:

> P199: "My view on automated systems haven't changed. I'm black so I'm already used to these kinds of things and having to explain to people why they should change their systems to be more inclusive of others or avoid making other uncomfortable."

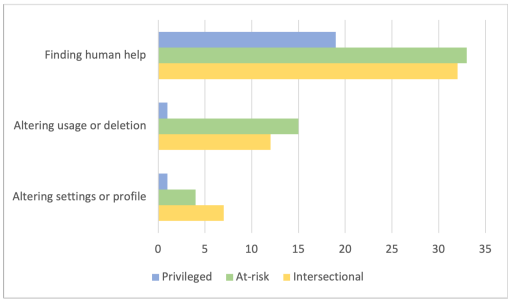


Fig. 3. Harm reduction strategies for different participant groups

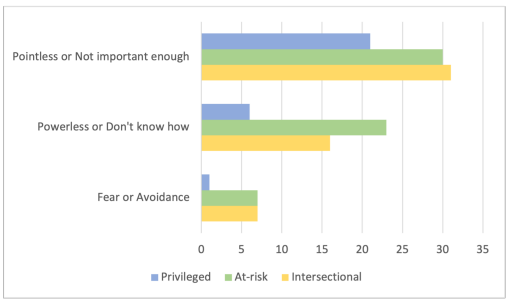


Fig. 4. Reasons for not acting for different participant groups

By far the most common response to being treated unfairly (see Figure 3) was to seek human help (65%, N=84), such as mentioned by P130 in Section 4.3.3. Participants repeatedly indicate that these human interventions have helped them to solve the issue. On the other end of the spectrum, there are participants who did not take action, and instead avoided the problem through compliance with an unfair system, through altering one's settings or profile. The strategy of altering settings is related to what Gasser has called Data Adjustment, a type of manual and ad-hoc workaround users engage in to solve immediate and pressing problems. In data adjustment, people "game" their computer systems by entering data that they knew were "inaccurate" or that did not reflect the spirit of the input data expected by the programs [47, 99]. Indeed, one intersectional participant notes:

> P203: "I began opting mostly to choose Male, because answering more honestly as "other" (or declining to answer) was costing me much needed income on an almost daily basis—and because Male is the gender I MOST closely identify with, is how all of the people I know interact with me, and is how I am generally perceived and treated other people out in public."

With regards to participants who noted to have not acted at all (see Figure 4), the most common response was that it would be pointless. This is summarised by one intersectional participant whose gender was not recognised by a shopping platform:

> P152: "Because it happens too often and life is too short."

Responses such as these signal the amount of relatively minor harms done by automated systems, many of which go unnoticed and unchecked. Further, even if the issue would be big enough to be raised, respondents note feeling powerless to do anything. Several intersectional participants mention being used to being treated unfairly due to personal attributes, mirroring the expected harm theme discussed above:

> P153: "Because i already take it as a price i must pay to be a black american, to be treated unfairly."

Finally, when it comes to systems required to gain opportunities, participants note not having responded due to fear of repercussions for doing so. This theme appears particularly often in hiring scenarios, and demonstrates the extent to which harms of compelled usage systems can go unnoticed:

> P122: "I am still in the interview process, and did not want to ruin my chances."

## 5 DISCUSSION

Our survey study in this paper provided participants with an opportunity to report on their experiences with unfair treatment by automated systems, to qualify those experiences as discriminatory, and to discuss what they did in response.

### 5.1 Perceptions of fairness

Our quantitative analysis of questionnaire responses found significant correlations between at-risk and intersectional characteristics and perceived unfair and discriminatory treatment. We found that intersectional participants are reporting unfairness and discrimination on particular platforms, such as registration systems and facial recognition systems. The former represents the many cases of representational harm we have counted, in which users are not recognised, excluded or unduly categorised by a system based on their (perceived) identity.

One of our main findings is that relatively minor cases of unfair treatment often go by without being noted or ameliorated. Many of the harms mentioned in our questionnaire are not considered grave enough to be responded to, or to be considered discriminatory, even when they involve protected characteristics. This emphasises that personal experiences and outcomes of unfair treatment are not exhausted by a framework of discrimination alone. The sense of fatigue and expected unfairness by some groups points to the need for better and simpler feedback mechanics in automated systems, in order to catch these incidents and raise the bar for everyday interactions with automated systems. This not only implies the need to place more 'humans in the loop', but to do so at points in the system interaction in which users feel essentialised or excluded, such as when registering for a service, or after a decision has been made about a user's eligibility.

The experiences of our intersectional participants surrounding automated systems can be broadly classified into three themes: essentialism, exclusion, and expectancy. Users "in between" normative categories are not recognised by systems, filed into adjacent classes they do not agree with, and excluded from opportunities, all of which lead to the anticipation of harm. We now address these themes in turn.

### 5.2 Addressing perceptions of unfair systems

*5.2.1 Essentialism.* Improving essentialist systems, which treat personal characteristics as fundamental or intrinsic to people, is notably difficult in representational scenarios. A possible goal can be to ensure equal representation of groups in a ranking, or to give due weight to different normative outlooks in classifiers that automate the enforcement of norms [125]. However, essentialism is not just a matter of creating better weighted or more granular categories: it also needs to be sought in revisionism, in the sense that the issues encountered by our participants can be ameliorated if particular personal attributes would be more easily mutable in a system. This is particularly important as many personal characteristics are increasingly considered as being in flux. Providing easier access to alter these features can significantly help to build trust in systems by users from at-risk backgrounds. In many cases, opening up these categories necessitates pushing back against cultural pressures and perceived needs of advertisers and institutions that require stable identity characteristics as predictors [16].

*5.2.2 Exclusion.* Experiences of exclusion, which are often distributive in kind, can be counteracted with transparency around inclusivity itself; that is, when a user is excluded from a system, more efforts should be made to explain to the user on the basis of which features the decision was (not) made. We found this particularly relevant for participants from the Global South, who partake in algorithmic microwork systems that cut across national lines, and are disadvantaged when compared to Global North participants — if they are able to partake at all. Our approach shows that

explanations ought to take place at the moment that users are confronted with a decision: many participants note that their frustration not only, or even primarily, stems from the decision itself, but from the lack of support or information they receive when wanting to enquire about it. This links to the issues of positionality faced by Explainable AI: beyond transparency of technical specifics, there is a need for a "relational transparency" [115] geared towards the kinds of explanations that users from different backgrounds and with different backgrounds require [74, 87, 88]. Such transparency needs to be built on robust usability studies, keeping in mind that, instead of devising two (or more) versions of the same software to serve different classes (genders, ethnicities, etc.), software requires inclusivity across the cognitive diversity that arises not only between or among different genders, but also within them. [119].

*5.2.3 Expectancy.* Preventing the expectancy of harms, whether distributive or representative, seems a crucial task for system engineers when designing for social justice. This expectancy is often connected by our participants to questions around and mistrust of system designers. This mirrors prior research highlighting difference in social groups when it comes to trusting human and computational systems [77]. As such, a blanket call for "more transparency" regarding system design is not enough. What is needed are explicit considerations for people who experience more negative encounters with automated systems. These considerations need to be baked into everyday algorithmic systems, such as microwork and recruitment platforms, and be part of the entire user journey—not just after the user is faced with a decision outcome. Prior work has suggested the inclusion of anti-discrimination clauses, including extensive explanations of how the programmers were committed to equity and anti-racism in their workplace and product [77]. Further, our participants' mistrust shows the need for users to clearly see why a system asks for particular characteristics. Clear feedback mechanisms after a decision outcome are crucial, and often lacking. The previously mentioned mistrust of human systems by vulnerable users means that beyond the addition of human support, the system itself should be able to give the user a footing in the reasons for outcomes that are unfavourable to them, and to provide useful advice on what they could possibly do next.

*5.2.4 Perception-based auditing.* Beyond implications for the design of algorithmic systems and their support structures, we believe that user studies of this kind can be the first step into a selection of algorithmic systems to be *audited*—that is, probing these systems by providing it with one or more inputs, while changing some attributes of that input, in order to identify situations and/or features that give rise to negative impacts [19, 83, 102]. An important aspect of audit studies is to specify how social problems manifest in technical systems [1, 6], and experiences by vulnerable individuals and communities can yield important cues in this context. It has been pointed out that everyday users of algorithmic systems are often best poised to detect and raise awareness about harmful behaviors that they encounter in the course of their everyday interactions with these systems [104]. Vulnerable individuals often have a lot of experience with particular systems—the gig work platforms discussed in this paper are no exception. For instance, the suspicion of being screened out of gig work opportunities due to gender identity as discussed by P203 in Section 4.2 could be further investigated by constructing several fictitious queries in response to job postings.

### 5.3 Limitations

The main limitation of this study arguably lies in our approach to the concept of intersectional harm, which in a sense is both not specific enough and too specific. On the one hand, there is space for more categorical complexity: after all, an analysis of intersectionality can include additional forms of multilevel, hierarchical, or contextual modeling, which can introduce more complexity in estimation and interpretation—for instance, by not simply asking about the effect of, e.g., ethnicity

on experiences of discrimination, but also how that effect differs for different gender identities [81]. Further, a larger empirical study that categorises the platforms mentioned by participants could employ these categorised platforms in a multilevel regression analysis using this context as a higher-order level of analysis. Such multilevel regression analysis has been argued to be more helpful in the context of intersectional research than multiple regression with interaction terms, as it explicitly accounts for "the social contexts of inequality by animating context itself as a unit of analysis and source of variance" [103, p.15].

At the same time, more can be done to explore the nature and extent of the inequalitites and differences in our data. There is a need to historicise discriminatory practices, as discrimination is not just experienced differently, but also shaped in a complex series of meaning that owes to past experiences [62]. This also means to reframe what we call *unfairness* here as a matter of *injustice* [13]. Analysing these processes means to critique the integrity of categorical distinctions themselves, and realise they are often themselves by-products of oppression [85]. We realise that our analysis of intersectional "categories" evades the problem of balancing the stability and fluidity of inequalities so they are sufficiently stable as to be available for empirical analysis, while recognizing that they change [120]. The question of how intersectional categories can be operationalised remains an open one.

A second limitation of this study lies in the selection of participants through Prolific, which led to many reported experiences of unfairness in the context of microwork and survey systems. Using other ways of recruiting might have altered the position of this theme in the top platforms, yet without necessarily invalidating what was reported for other platforms. We also note that our survey, like any other self-report measure, necessarily is subject to a number of social and cognitive limitations regarding perception, interpretation, and disclosure [75].

Finally, and from a methodological perspective, there are different ways to group thematic concepts. While we report strong agreement rates in Section 3.3, we also need to emphasise the importance of perspective as emphasised in standpoint theory with regards to our own analysis [59, 90]. That is, we must acknowledge that our perspectives and backgrounds influence our categories and findings [17]. It should be emphasised that the goal of this paper was not to determine the 'best' approach to evaluate fairness perceptions; rather, we aimed to offer a space for different individuals to share what they consider to be noteworthy experiences. We believe that the Critical Incident Technique offers a valuable tool for human-centered computing in this regard.

## 6 CONCLUSION

Automated computational systems can cause harm to their users in many different ways, both in and outside a context of discrimination. In computer science, approaches to fairness typically focus on ameliorating these issues by focusing on data preparation, model-learning or post-processing. However, there is a lack of interventions that attempt to "more profoundly, deeply, and adeptly interrogate the power structures and issues that undergird these critical narratives" when it comes to automated computation system's harms and risks [40, p.177]. In this paper, we have focused on marginalised participants narratives on unfair treatment, in order to underscore that the interaction between humans and automated computational systems matters, even when discussing relatively minor incidents, and that platforms that are unfair for one group will privilege another. Allowing users from different backgrounds to recount their experiences *in situ*, we aimed to explore how particular features relate to experiences and understandings of unfairness. Our critical incident approach also has underscored that explicit discrimination is only a part of this fairness discussion [55]. Beyond the results of disparate impact, a sense of not being heard or seen by systems that make decisions about opportunities or access is pervasive in our data.

Rather than viewing such minor incidents as trivial, system engineers need to incorporate more means of human interaction throughout the user journey. Preventing further public backlash against computational, particularly AI-based, systems [40, p.165] involves the need for what we might call *localised transparency* about the inclusion of features and the background of designers. This can help address the general sense, returning in our data on different occasions, of not being helped when an unfair situation arises. Dealing with these experiences can simply not be fully automated; the human interaction springing from it needs to take one step in the uncomfortable process of understanding that increasing the status of marginalised people often means to address how privileges of privileged or powerful groups will change as well [65]. Intersectionality also points out a need to move beyond notions of fairness and towards those of justice: this means to understand that unfairness does not take place in separate, singular systems but is part of the overarching social structures we partake in [13]. This significantly impacts the research process as well: we believe studies on perceptions of fairness should always be done with positionality in mind [68], and should not be undertaken without meaningful collaboration between researchers from both technical and critical disciplines, and from diverse personal backgrounds.

To conclude, the recurring stories about registration and microwork systems by our participants have provided us with an unexpected salutary lesson. At first glance, such systems might seem odd candidates to top an unfairness list. But when one considers how they foreground self-categorisation, it quickly becomes apparent that they provide the opportunity for critical reflection [16]. When negative causal effects are experienced in such a temporally proximate and linear fashion — e.g., an individual is disqualified from participation due to non-binary gender status — it not only sharpens perceptions of injustice, but offers an opportunity to rethink user feedback. Registration and microwork systems stand in contrast to many automated computational systems that are once-removed from users, wherein the machinations of disparate impact are more causally delayed, or embedded and distributed across complex systems, muting perceptual acuity. What the reported experiences in this survey show is that automated computational systems require open communications with users *while* unfairness is being experienced.

## ACKNOWLEDGMENTS

We thank the anonymous reviewers. This work was supported by EPSRC under grant EP/R033188/1. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising. The data supporting this article has been deposited in The King's Open Research Data System (KORDS) at 10.18742/20499216. It is not openly available due to data handling and confidentiality agreements and may be shared on request through completing a data access agreement.

## REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 252–260. https://doi.org/10.1145/3351095.3372871 arXiv:1912.04883

[2] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society* 20, 3 (2018), 973–989. https://doi.org/10.1177/1461444816676645

[3] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI and Society* 35, 3 (2020), 611–623. https://doi.org/10.1007/s00146-019-00931-w

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. https://doi.org/10.1016/j.inffus.2019.12.012 arXiv:1910.10045

[5] Carl Auerbach and Louise B. Silverstein. 2003. *Qualitative data: An introduction to coding and analysis.* New York University Press, New York & London. 1–202 pages. https://doi.org/10.5860/choice.41-4324

[6] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. (2021), 1–34. arXiv:2102.04256 http://arxiv.org/abs/2102.04256

[7] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: Reorienting the Study of Algorithmic Fairness around Issues of Power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/3351095.3372859

[8] Solon Barocas and Andrew Selbst. 2016. Big Data's Disparate Impact. *California law review* 104, 1 (2016), 671–729. https://doi.org/10.15779/Z38BG31

[9] Greta R. Bauer, Siobhan M. Churchill, Mayuri Mahendran, Chantel Walwyn, Daniel Lizotte, and Alma Angelica Villa-Rueda. 2021. Intersectionality in quantitative research: A systematic review of its emergence and applications of theory and methods. *SSM - Population Health* 14, February (2021), 100798. https://doi.org/10.1016/j.ssmph.2021.100798

[10] David Beer. 2009. Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society* 11, 6 (2009), 985–1002. https://doi.org/10.1177/1461444809336551

[11] Ruha Benjamin. 2019. *Race after Technology.* Polity, Cambridge.

[12] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[13] Cynthia L. Bennett and Os Keyes. 2019. What is the Point of Fairness? Disability, AI and The Complexity of Justice. (2019). https://doi.org/10.1145/3386296.3386301 arXiv:1908.01024

[14] Nora Berenstain. 2016. Epistemic Exploitation. *Ergo, an Open Access Journal of Philosophy* 3, 20201214 (2016), 569–590. https://doi.org/10.3998/ergo.12405314.0003.022

[15] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Conference on Fairness, Accountability and Transparency*. PMLR, New York, NY, 149–159. arXiv:1712.03586 http://arxiv.org/abs/1712.03586

[16] Rena Bivens and Oliver L. Haimson. 2016. Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society* 2, 4 (2016), 2056305116672486. https://doi.org/10.1177/2056305116672486 arXiv:https://doi.org/10.1177/2056305116672486

[17] Alan Borning and Michael Muller. 2012. Next Steps for Value Sensitive Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1125–1134. https://doi.org/10.1145/2207676.2208560

[18] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society* 15, 5 (2012), 662–679. https://doi.org/10.1080/1369118X.2012.678878

[19] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data and Society* 8, 1 (2021). https://doi.org/10.1177/2053951720983865

[20] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[21] Lee D. Butterfield, William A. Borgen, Norman E. Amundson, and Asa Sophia T. Maglio. 2005. Fifty years of the critical incident technique: 1954-2004 and beyond. *Qualitative Research* 5, 4 (2005), 475–497. https://doi.org/10.1177/1468794105056924

[22] Devon W. Carbado. 2013. Colorblind Intersectionality. *Signs* 38, 4 (2013), 811–845.

[23] Logan S. Casey, Sari L. Reisner, Mary G. Findling, Robert J. Blendon, John M. Benson, Justin M. Sayde, and Carolyn Miller. 2019. Discrimination in the United States: Experiences of lesbian, gay, bisexual, transgender, and queer Americans. *Health Services Research* 54, S2 (2019), 1454–1466. https://doi.org/10.1111/1475-6773.13229

[24] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis.* Sage, London, Thousands Oaks and New Delhi. https://doi.org/10.1016/j.lisr.2007.11.003 arXiv:arXiv:1011.1669v3

[25] Arun Chinmayi. 2020. AI and the Global South: Designing for Other Worlds. In *Oxford Handbook of Ethics in AI.* Oxford University Press, New York, 229–237. https://ssrn.com/abstract=3403010

[26] Sumi Cho, Kimberlé Williams Crenshaw, and Leslie McCall. 2013. Toward a Field of Intersectionality Studies: Theory, Applications, and Praxis. *Signs* 38, 4 (2013), 785–810.

[27] Julie E. Cohen. 2019. *Between truth and power.* Oxford University Press, New York.

[28] Patricia Hill Collins. 2019. *Intersectionality as Critical Social Theory.* Duke University Press, Durham & London. https://doi.org/10.1177/0891243220973228

[29] Jason A. Colquitt and Jessica B. Rodell. 2015. Measuring Justice and Fairness. In *The Oxford Handbook of Justice in the Workplace*, Russell S. Cropanzano and Maureen L. Ambrose (Eds.). Oxford University Press, 187–202. https:

//doi.org/10.1093/oxfordhb/9780199981410.013.8
[30] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 1, 8 (1989). https://doi.org/10.3917/drs1.108.0465
[31] Natalia Criado, Xavier Ferrer, and Jose Such. 2021. Attesting Digital Discrimination Using Norms. *International Journal of Interactive Multimedia and Artificial Intelligence* 6, 5 (2021), 16. https://doi.org/10.9781/ijimai.2021.02.008
[32] Natalia Criado and Jose Such. 2019. Digital Discrimination. In *Algorithmic Regulation*. Oxford University Press, 82–97.
[33] Catherine Crisp. 2014. White and Lesbian: Intersections of Privilege and Oppression. *Journal of Lesbian Studies* 18, 2 (2014), 106–117. https://doi.org/10.1080/10894160.2014.849161
[34] John Danaher. 2020. Sexuality. In *The Oxford Handbook of Ethics of AI*. Oxford University Press, New York, Chapter 20, 403–417.
[35] Kathy Davis. 2008. Intersectionality as buzzword: A sociology of science perspective on what makes a feminist theory successful. *Feminist Theory* 9, 1 (2008), 67–85. https://doi.org/10.1177/1464700108086364 arXiv:Davis2008
[36] Michael A. De Vito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms ruin everything": #RIPTwitter, folk theories, and resistance to algorithmic change in social media. *Conference on Human Factors in Computing Systems - Proceedings* 2017-May (2017), 3163–3174. https://doi.org/10.1145/3025453.3025659
[37] Morton Deutsch, Peter T. Coleman, and Eric C. Marcus. 2006. *The Handbook of Conflict Resolution: Theory and Practice*. Jossey-Bass, San Fransisco. 310 pages. http://books.google.com.au/books?hl=en{&}lr={&}id=rw61VDID7U4C{&}oi=fnd{&}pg=PR7{&}dq=professional+conflict+management{&}ots=zblqs8ptTt{&}sig=PsDHja6pwPcuwwBLeFNGmLL4fo8{#}v=onepage{&}q=interpersonal{&}f=false
[38] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data feminism*. MIT Press, Cambridge & London. https://doi.org/10.1080/1369118x.2020.1836249
[39] Angel P Diz, Antonio Carvajal-Rodríguez, and David OF Skibinski. 2011. Multiple hypothesis testing in proteomics: a strategy for experimental work. *Molecular & Cellular Proteomics* 10, 3 (2011).
[40] Marcus D. Dubber, Frank Pasquale, and Sunit Das. 2020. *The Oxford Handbook of Ethics of AI*. Oxford University Press, New York.
[41] Viriginia Eubanks. 2018. *Automating Inequality*. St. Martin's Press, New York.
[42] Xavier Ferrer, Tom van Nuenen, Jose Such, Mark Coté, and Natalia Criado. 2021. Bias and Discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society Magazine* 40, 2 (2021), 72–80.
[43] Xavi Ferrer Aran, Tom Van Nuenen, Natalia Criado, and Jose Such. 2021. Discovering and Interpreting Biased Concepts in Online Communities. *IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1. https://doi.org/10.1109/TKDE.2021.3139680
[44] John Flanagan. 1954. The critical incident technique. *Psychological bulletin* 59, 4 (1954), 257–72. http://www.ncbi.nlm.nih.gov/pubmed/19586159
[45] Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, Oxford and New York. https://doi.org/10.1177/002114006803500105
[46] Clare Garvie, Alvaro M. Bedoya, and Jonathan Frankle. 2016. *The Perpetual Line-Up*. Technical Report. Georgetown Law Center on Privacy & Technology, Washington DC.
[47] Les Gasser. 1986. The Integration of Computing and Routine Work. *ACM Transactions on Information Systems (TOIS)* 4, 3 (1986), 205–225. https://doi.org/10.1145/214427.214429
[48] Timnit Gebru. 2020. Race and gender. In *The Oxford handbook of ethics of aI*. Oxford University Press, Oxford, 251–269.
[49] Janneke Gerards and Raphaële Xenidis. 2021. *Algorithmic discrimination in Europe*. Publications Office of the European Union,, Luxembourg. 192 pages.
[50] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019), 2221–2231. https://doi.org/10.1145/3292500.3330691 arXiv:1905.01989
[51] Tarleton Gillespie. 2012. The relevance of algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot (Eds.). MIT press, Cambridge, 167–194.
[52] Barney G Glaser and Anselm L Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Transaction Publishers, New Brunswick and London. https://doi.org/10.2307/2575405 arXiv:9809069v1 [arXiv:gr-qc]
[53] Evelyn Nakano Glenn. 2002. *Unequal Freedom: How Race and Gender Shaped American Citizenship and Labor*. Harvard University Press, Cambridge & London.
[54] João Gonçalves, Joana Malta-Vacas, Monette Louis, Laurent Brault, Denyse Bagrel, Carolino Monteiro, and Miguel Brito. 2005. Modulation of translation factor's gene expression by histone deacetylase inhibitors in breast cancer cells. *Clinical Chemistry and Laboratory Medicine* 43, 2 (2005), 151–156. https://doi.org/10.1515/CCLM.2005.025

[55] Nina Grgić-Hlača, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *WWW 2018: The 2018 Web Conference*. Lyon, France, 903–912. https://doi.org/10.1145/3178876.3186138 arXiv:1802.09548

[56] Ange-Marie Hancock. 2019. Empirical Intersectionality: A Tale of Two Approaches. *The Palgrave Handbook of Intersectionality in Public Policy* 3, 2 (2019), 95–132. https://doi.org/10.1007/978-3-319-98473-5_5

[57] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 501–512. https://doi.org/10.1145/3351095.3372826 arXiv:1912.03593

[58] Rita Hardiman, Bailey W. Jackson, and Pat Griffin. 2007. Conceptual Foundations for Social Justice Education. In *Teaching for diversity and social justice*, M. Adams, L. A. Bell, and P. Griffin (Eds.). Routledge, 35–66.

[59] Sandra Harding. 2004. *The Feminist Standpoint Theory Reader*. Routledge, New York & London.

[60] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 392–402. https://doi.org/10.1145/3351095.3372831

[61] Hoda Heidari, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. 2018. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. (2018), 1–15. https://doi.org/10.48550/arXiv.1806.04959

[62] Kerry Hendricks, Nick Deal, Albert J. Mills, and Jean Helms Mills. 2021. Intersectionality as a matter of time. *Management Decision* 59, 11 (2021), 2567–2582. https://doi.org/10.1108/MD-02-2019-0264

[63] Julian PT Higgings and Jonathan J Deeks. 2008. *Selecting studies and Collecting Data*. John Wiley & Sons Ltd. 151–185 pages.

[64] Sophie Hilgard and Himabindu Lakkaraju. 2021. Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring. (2021). arXiv:2012.00423v2 www.aaai.org

[65] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information Communication and Society* 22, 7 (2019), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

[66] Katja Hujo. 2021. Social protection and inequality in the global South: Politics, actors and institutions. *Critical Social Policy* 41, 3 (2021), 343–363. https://doi.org/10.1177/02610183211009899

[67] Lilly Irani. 2015. The cultural work of microwork. *New Media & Society* 17, 5 (2015), 720–739. https://doi.org/10.1177/1461444813511926

[68] Alison Jagger. 2008. *Just Methods: An Interdisciplinary Feminist Reader*. Paradigm Pubishers, Boulder & London.

[69] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021). https://doi.org/10.1145/3476046

[70] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520

[71] Luisa Kcomt. 2019. Profound health-care discrimination experienced by transgender people: rapid systematic review. *Social Work in Health Care* 58, 2 (2019), 201–219. https://doi.org/10.1080/00981389.2018.1532941

[72] Yujin Kim, Jennifer Dykema, John Stevenson, Penny Black, and D. Paul Moberg. 2019. Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys. *Social Science Computer Review* 37, 2 (2019), 214–233. https://doi.org/10.1177/0894439317752406

[73] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. *Leibniz International Proceedings in Informatics, LIPIcs* 67 (2017), 1–23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43 arXiv:arXiv:1609.05807v2

[74] Hana Kopecka and Jose Such. 2020. Explainable AI for Cultural Minds. In *Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction*. https://sites.google.com/view/dexahai-at-ecai2020/home Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction , DEXAHAI ; Conference date: 07-09-2020.

[75] Nancy Krieger, Kevin Smith, Deepa Naishadham, Cathy Hartman, and Elizabeth M. Barbeau. 2005. Experiences of discrimination: Validity and reliability of a self-report measure for population health research on racism and health. *Social Science and Medicine* 61, 7 (2005), 1576–1596. https://doi.org/10.1016/j.socscimed.2005.03.006

[76] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs Discussion-Based Social Division. In *20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017)*. Portland, OR, 1035–1048.

[77] Min Kyung Lee and Kate Rich. 2021. Who is included in human perceptions of ai?: Trust and perceived fairness around healthcare ai and cultural mistrust. In *CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, Yokohama. https://doi.org/10.1145/3411764.3445570

[78] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology* 31, 4 (2018), 611–627. https://doi.org/10.1007/s13347-017-0279-x

[79] Deborah Lupton. 2018. How do data come to matter? Living and becoming with personal data. *Big Data and Society* 5, 2 (2018), 1–11. https://doi.org/10.1177/2053951718786314

[80] Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. 2020. Implications of AI (Un-)Fairness in Higher Education Admissions: The Effects of Perceived AI (Un-)Fairness on Exit, Voice and Organizational Reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 122–130. https://doi.org/10.1145/3351095.3372867

[81] Leslie McCall. 2005. The Complexity of Intersectionality. *Signs* 30, 3 (2005), 1771–1800.

[82] Nora McDonald and Shimei Pan. 2020. Intersectional AI: A Study of How Information Science Students Think about Ethics and Their Impact. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020). https://doi.org/10.1145/3415218

[83] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. *Auditing Algorithms: Understanding Algorithmic Systems from the Outside In.* now Publishers, Hanover & Delft.

[84] Stefania Milan and Emiliano Treré. 2019. Big Data from the South(s): Beyond Data Universalism. *Television and New Media* 20, 4 (2019), 319–335. https://doi.org/10.1177/1527476419837739

[85] Joya Misra, Celeste Vaughan Curington, and Venus Mary Green. 2020. Methods of intersectional research. *Sociological Spectrum* 0, 0 (2020), 1–20. https://doi.org/10.1080/02732173.2020.1791772

[86] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 205395171667967. https://doi.org/10.1177/2053951716679679

[87] Francesca Mosca and Jose Such. 2022. An explainable assistant for multiuser privacy. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 1–45.

[88] Francesca Mosca and Jose M. Such. 2021. ELVIRA: An Explainable Agent for Value and Utility-Driven Multiuser Privacy. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 916–924.

[89] Arvind Narayanan. 2018. 21 fairness definitions and their politics. In *Tutorial at FAT Conference 2018.*

[90] Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–16. https://doi.org/10.1145/3313831.3376392

[91] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishers, New York. https://doi.org/10.5860/crl.78.3.403

[92] Richard Owen, Phil Macnaghten, and Jack Stilgoe. 2012. Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy* 39, 6 (2012), 751–760. https://doi.org/10.1093/scipol/scs093

[93] Leonard J Paas and Meike Morren. 2018. Please do not answer if you are reading this: Respondent attention in online panels. *Marketing Letters* 29, 1 (2018), 13–21.

[94] Pega. 2018. *What Consumers Really Think About AI : A Global Study Executive summary : The AI / consumer paradox.* Technical Report. Pega, Cambridge, MA. https://www.pega.com/ai-survey

[95] Jennifer Pierre, Roderic Crooks, Morgan E. Currie, Britt S. Paris, and Irene V. Pasquetto. 2021. Getting ourselves together: Data-centered participatory design research and epistemic burden. *Conference on Human Factors in Computing Systems - Proceedings* (may 2021). https://doi.org/10.1145/3411764.3445103

[96] Emma Pierson. 2018. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124* (2018). arXiv:1712.09124v2 https://github.com/epierson9/algorithmic{_}

[97] Emilee Rader, Rick Wash, and Brandon Brooks. 2012. Stories as informal lessons about security. *SOUPS 2012 - Proceedings of the 8th Symposium on Usable Privacy and Security* (2012). https://doi.org/10.1145/2335356.2335364

[98] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20 (2018), 5–14. https://doi.org/10.1007/s10676-017-9430-8

[99] Kopo Ramokapane, Gaurav Misra, Jose Such, and Sören Preibusch. 2021. Truth or Dare: Understanding and Predicting How Users Lie and Provide Untruthful Data Online. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[100] Amanda Rodriguez, Anette Agardh, and Benedict Oppong Asamoah. 2018. Self-Reported Discrimination in Health-Care Settings Based on Recognizability as Transgender: A Cross-Sectional Study Among Transgender U.S. Citizens. *Archives of Sexual Behavior* 47, 4 (2018), 973–985. https://doi.org/10.1007/s10508-017-1028-z

[101] Johnny Saldana. 2013. *The Coding Manual for Qualitative Researchers.* Sage, London.

[102] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. 1–24. https://fdocuments.in/document/auditing-algorithms-research-methods-for-detecting-discrimination-.html

[103] Nicholas A. Scott and Janet Siltanen. 2017. Intersectionality and quantitative methods: assessing regression from a feminist perspective. *International Journal of Social Research Methodology* 20, 4 (2017), 373–385. https://doi.org/10.1080/13645579.2016.1201328

[104] Hong Shen, Alicia Devos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021). https://doi.org/10.1145/3479577 arXiv:2105.02980

[105] Donghee Shin. 2020. User Perceptions of Algorithmic Decisions in the Personalized AI System:Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability. *Journal of Broadcasting and Electronic Media* 64, 4 (2020), 541–565. https://doi.org/10.1080/08838151.2020.1843357

[106] Mona Sloane. 2019. Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice. In *Proceedings of the Weizenbaum Conference 2019*.

[107] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2021. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. *arXiv preprint arXiv:2103.12016* (2021).

[108] Jose Such. 2017. Privacy and autonomous systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*. 4761–4767.

[109] Jose Such, Agustín Espinosa, and Ana García-Fornes. 2014. A survey of privacy in multi-agent systems. *The Knowledge Engineering Review* 29, 3 (2014), 314–344.

[110] Jose Such, Joel Porter, Sören Preibusch, and Adam Joinson. 2017. Photo Privacy Conflicts in Social Media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, 3821–3832. https://doi.org/10.1145/3025453.3025668

[111] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Ssrn* (2013). https://doi.org/10.2139/ssrn.2208240 arXiv:1301.6822

[112] Jakita O. Thomas, Nicole Joseph, Arian Williams, Chanrtel Crum, and Jamika Burge. 2018. Speaking Truth to Power: Exploring the Intersectional Experiences of Black Women in Computing. *2018 Research on Equity and Sustained Participation in Engineering, Computing, and Technology, RESPECT 2018 - Conference Proceedings* April 2019 (2018). https://doi.org/10.1109/RESPECT.2018.8491718

[113] Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. 2019. Considerations for AI fairness for people with disabilities. *AI Matters* 5, 3 (2019), 40–63. https://doi.org/10.1145/3362077.3362086

[114] United Nations. 2020. *World Economic Situation and Prospects*. Technical Report. 163–171 pages. https://doi.org/10.18356/036ade46-en

[115] Tom Van Nuenen, Xavier Ferrer, Jose Such, and Mark Cote. 2020. Transparency for Whom? Assessing Discriminatory Artificial Intelligence. *Computer* 53, 11 (2020), 36–44. https://doi.org/10.1109/MC.2020.3002181

[116] Kami Vaniea and Yasmeen Rashidi. 2016. Tales of software updates: The process of updating software. *Conference on Human Factors in Computing Systems - Proceedings* (2016), 3215–3226. https://doi.org/10.1145/2858036.2858303

[117] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *ACM/IEEE International Workshop on Software Fairness (FairWare)*. IEEE, 1–7. https://doi.org/10.1145/3194770.3194776

[118] Roderik F. Viergever. 2019. The Critical Incident Technique: Method or Methodology? *Qualitative Health Research* 29, 7 (2019), 1065–1079. https://doi.org/10.1177/1049732318813112

[119] Mihaela Vorvoreanu, Lingyi Zhang, Yun Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, and Margaret Burnett. 2019. From gender biases to gender-inclusive design: An empirical investigation. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–14. https://doi.org/10.1145/3290605.3300283

[120] Sylvia Walby, Jo Armstrong, and Sofia Strid. 2012. Intersectionality: Multiple inequalities in social theory. *Sociology* 46, 2 (2012), 224–240. https://doi.org/10.1177/0038038511416164

[121] John Wihbey. 2015. The possibilities of digital discrimination: Research on e-commerce, algorithms and big data. *Journalist's resource* (2015).

[122] David R. Williams, Hector M. Gonzalez, Stacey Williams, Selina A. Mohammed, Hashim Moomal, and Dan J. Stein. 2008. Perceived discrimination, race and health in South Africa. *Social Science and Medicine* 67, 3 (2008), 441–452. https://doi.org/10.1016/j.socscimed.2008.03.021

[123] Pak Hang Wong. 2020. Democratizing Algorithmic Fairness. *Philosophy and Technology* 33, 2 (2020), 225–244. https://doi.org/10.1007/s13347-019-00355-w

[124] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeff Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April

(2018), 1–14. https://doi.org/10.1145/3173574.3174230

[125] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A fair top-k ranking algorithm. *International Conference on Information and Knowledge Management, Proceedings* Part F1318 (2017), 1569–1578. https://doi.org/10.1145/3132847.3132938 arXiv:1706.06368

[126] Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 4 (2017), 1060–1089. https://doi.org/10.1007/s10618-017-0506-1

# A APPENDIX

## A.1 Unfair treatment by automated computational systems - questionnaire

*A.1.1 Instructions.* In this survey, we are interested in cases where you felt to be treated unfairly, unethically, or differently, based on who you are, by an automated system such as an online service, computer, application, or smart device. With "who you are", we refer to personal details (gender, age, sex, religion, disability, etc.) or social groups or communities you consider yourself to be a part of. Please note that we are NOT referring to other people treating you unfairly online, but instead the output of a system, computer, or device. Note: Please finish the entire questionnaire, after which you will be redirected to Prolific.

*A.1.2 Experiences of unfair treatment.*

(1) What is your Prolific ID? [Open-ended]
(2) When was the last time you felt to be treated unfairly when interacting with an automated computational system? Recall: we are talking about an automated system (NOT another person) treating you unfairly because of your personal details.
   - Less than a week ago (1)
   - Less than a month ago (2)
   - Less than a year ago (3)
   - Less than two years ago (4)
   - More than two years ago (5)
   - Never (6)
(3) What was the automated system or (online) platform you were using when you felt treated unfairly? [Open-ended] Recall: we are talking about an automated system (NOT another person) treating you unfairly because of your personal details.
(4) Why did you feel you were treated unfairly? (open-ended) Recall: we are talking about an automated system (NOT another person) treating you unfairly because of your personal details.
(5) Could you explain what happened? Please be as detailed as possible. [Open-ended]
(6) On a scale from 1 to 100 how much mental pain, hurt, or distress did the experience cause you? Note: 1 means very little mental pain, hurt, or distress; 100 means a lot of mental pain, hurt, or distress.
(7) Did you take any action following the experience? For instance, contact the owner of the system or platform, delete your profile, change your privacy settings, and so on.
   - Yes (1)
   - No (2)
(8) What did you do and why? [Open-ended, shown if answer previous question is "Yes"]
(9) Why did you not take action? [Open-ended, shown if answer previous question is "No"]
(10) How many zebras do you have in your pocket? [Attention check]
   - One (1)
   - Two (2)
   - More than two (3)

- None (4)

(11) Have you decreased your usage of the automated system that you had the negative experience with?
- Not decreased at all (1)
- Slightly decreased (2)
- Moderately decreased (3)
- Considerably decreased (4)
- Stopped using it entirely (5)

(12) How have you changed your view on the automated system you had the negative experience with? [Open-ended]

(13) Do you feel that your unfair treatment by the online system, computer, or smart device was discriminatory? Please explain your answer. [Open-ended]

*A.1.3 Technical literacy.* The next questions are about your perceived ability to engage with computers, IT services, and the Internet.

(1) How long have you been using computers?
- Less than a year (1)
- 1 - 5 years (2)
- 5 - 10 years (3)
- Over 10 years / all my life (4)

(2) How would you rate your own internet literacy?
- Very good (1)
- Good (2)
- Acceptable (3)
- Poor (4)
- Very poor (5)
- Don't know (6)

(3) Please indicate the extent to which you agree or disagree with the following statements [Likert 1-5]
- I feel comfortable using digital technologies.
- I am aware of various types of digital technologies.
- I am willing to learn more about digital technologies.
- I feel threatened when others talk about digital technologies.
- I feel that I am behind my peers in using digital technologies.
- I think that it is important for me to improve my digital fluency.

*A.1.4 Demographic questions .* Finally, we have some questions about your personal details.

(1) How would you describe your ethnicity?
- Black or African American (1)
- American Indian or Alaska Native (2)
- Asian (3)
- White (4)
- Latinx (5)
- Native Hawaiian or Pacific Islander (6)
- Mixed (7)
- Other, namely (open-ended) (8)
- Prefer not to say (9)

(2) How would you describe your sexual orientation? Note: multiple answers are possible.

- Gay (1)
- Lesbian (2)
- Bisexual (3)
- Queer (8)
- Straight (heterosexual) (4)
- Questioning or unsure (5)
- Other, namely (6) (open-ended)
- Prefer not to say (7)

(3) What gender do you identify with?
- Female (1)
- Male (2)
- Non-binary (3)
- Gender neutral (4)
- Fluid (5)
- Transgender (6)
- Agender (7)
- Pangender (8)
- Questioning or unsure (9)
- Other, namely (10) (open-ended)
- Prefer not to say (11)

(4) What is your age?
- Under 18 years old (1)
- 18-24 years old (2)
- 25-34 years old (3)
- 35-44 years old (4)
- 45-55 years old (5)
- 55-64 years old (6)
- 65-74 years old (7)
- 75 or older (8)
- Prefer not to say (9)

(5) Do you have a disability?
- Yes (1)
- Questioning or unsure (2)
- No (3)
- Prefer not to say (4)

(6) Do you have children?
- Yes (1)
- No (2)
- Prefer not to say (3)

(7) Do you consider yourself religious?
- Yes (1)
- No (2)
- Questioning or unsure (3)
- Prefer not to say (4)

(8) What is your religion?
- Muslim (1)
- Mormon (2)
- Orthodox church (e.g. Greek or Russian Orthodox Church) (3)

- Jewish (4)
- Christian (11)
- Protestant (5)
- Christian Scientist (6)
- Seventh-day Adventist (7)
- Roman Catholic (8)
- Other, namely (9) (open-ended)
- Prefer not to say (10)

(9) What is the highest level of school you have completed or the highest degree you have received?
- Less than high school degree (1)
- High school graduate (high school diploma or equivalent including GED) (2)
- Some college but no degree (3)
- Associate degree in college (2-year) (4)
- Bachelor's degree in college (4-year) (5)
- Master's degree (6)
- Doctoral degree (7)
- Professional degree (JD, MD) (8)
- Prefer not to say (9)

(10) Where are you located? (open-ended)

(11) What is your marital status?
- Single (never married) (1)
- Married (2)
- In a domestic partnership (3)
- Divorced (4)
- Widowed (5)
- Prefer not to say (6)

(12) Information about income is very important to understand. Would you please give your best guess? This includes your entire household income in 2020 before taxes.
- Less than $10,000 (1)
- $10,000 to $19,999 (2)
- $20,000 to $29,999 (3)
- $30,000 to $39,999 (4)
- $40,000 to $49,999 (5)
- $50,000 to $59,999 (6)
- $60,000 to $69,999 (7)
- $70,000 to $79,999 (8)
- $80,000 to $89,999 (9)
- $90,000 to $99,999 (10)
- $100,000 to $149,999 (11)
- $150,000 or more (12)
- Prefer not to say (13)

(13) Finally, do you have any feedback you would like to give us regarding the questionnaire? For instance, were certain questions difficult to answer? Was one part more confusing than others? [Open-ended]

Received January 2022; revised April 2022; accepted May 2022