Proceedings of the Twelfth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2024)

Investigating What Factors Influence Users' Rating of Harmful Algorithmic Bias and Discrimination

Sara Kingsley[1*], Jiayin Zhi[1*], Wesley Hanwen Deng[1], Jaimie Lee[1], Sizhe Zhang[1], Motahhare Eslami[1†], Kenneth Holstein[1†], Jason I. Hong[1†], Tianshi Li[2†], Hong Shen[1†]

[1] Carnegie Mellon University
[2] Northeastern University

Abstract

There has been growing recognition of the crucial role users, especially those from marginalized groups, play in uncovering harmful algorithmic biases. However, it remains unclear how users' identities and experiences might impact their rating of harmful biases. We present an online experiment (N=2,197) examining these factors: demographics, discrimination experiences, and social and technical knowledge. Participants were shown examples of image search results, including ones that previous literature has identified as biased against marginalized racial, gender, or sexual orientation groups. We found participants from marginalized gender or sexual orientation groups were more likely to rate the examples as more severely harmful. Belonging to marginalized races did not have a similar pattern. Additional factors affecting users' ratings included discrimination experiences, and having friends or family belonging to marginalized demographics. A qualitative analysis offers insights into users' bias recognition, and why they see biases the way they do. We provide guidance for designing future methods to support effective user-driven auditing.

1 Introduction

Over the past decade, a wide range of biased and harmful behaviors have been documented in algorithmic systems, which disproportionately impact marginalized groups and raise significant concerns regarding fairness, accountability, and transparency in AI (Eubanks 2018; Noble 2018). Past research has developed methods to help experts audit algorithmic systems for harmful behaviors (Metaxa et al. 2021; Costanza-Chock, Raji, and Buolamwini 2022; Shen et al. 2021). These expert-driven audits have been successful in detecting many machine biases, but also suffer from many limitations, such as expert blindspots ("unknown unknowns") and algorithmic behaviors that only emerge when a system is deployed with actual users (Shen et al. 2021).

Recent years have seen the rise of *user-driven* audits, which have potential to overcome some of these limitations. Here, end-users organically come together to collectively uncover and make sense of potentially harmful machine behaviors they encounter in their everyday lives (Shen et al. 2021; DeVos et al. 2022; Deng et al. 2023; Li et al. 2023). Members of marginalized groups, who are often disproportionately harmed by algorithmic biases, have worked collectively to challenge these biases in AI systems. For example, online content creators who believed they were demonetized for producing LGBTQ+ educational content (Bensinger and Albergotti 2019) or content about Black culture (Albergotti 2020) worked together to conduct A/B tests to evaluate what seemingly triggers a platform's algorithm to filter their content (Kingsley et al. 2022). The power of everyday users in surfacing harmful algorithmic behaviors has also motivated researchers to develop tools and methods to conduct *user-engaged* algorithm auditing, in which AI developers actively engage with end-users in auditing harmful algorithmic behaviors that might otherwise go undetected (Lam et al. 2022; Deng et al. 2023; Kiela et al. 2021; Ochigame and Ye 2021; OpenAI 2022; Cattell, Chowdhury, and Carson 2023).

While user-driven and -engaged algorithm audits have significant potential, we have limited understanding of how end-users' *identities* and *experiences* might impact their likelihood of perceiving and rating harmful algorithmic behaviors. Prior research in AI and HCI has investigated how demographic features might impact people's perceptions of bias, fairness, and harms in algorithmic systems (Kumar et al. 2021; Sap et al. 2019; Wang, Harper, and Zhu 2020; Jiang et al. 2021) and has offered initial qualitative evidence of the impact of user identity and experience on perceptions of algorithmic behaviors (DeVos et al. 2022). However, many questions remain. For example, are users who are directly affected by and have lived experiences with societal discrimination (e.g., people of color, women, LGBTQ+, etc.) more likely of rating related algorithmic biases as more harmful of detecting related algorithmic biases? How effective are other groups in the same biases? Developing a better understanding of these questions can provide useful design guidelines for supporting user-driven and -engaged auditing. In addition to providing insight into how to more effectively engage and organize users in the auditing process (e.g., who needs to be included?), as Deng et al. highlighted (Deng et al. 2023), identifying demographics and factors that are effective in rating harmful algorithmic bias can help us understand how we can most effectively recruit allies as auditors, to reduce the burden placed on marginalized end-

[*] These authors contributed equally.
[†] These authors contributed equally.

users. Such an effort can also improve user-driven auditing by broadening the pool of potential participants, thereby making recruiting easier without lessening audit quality.

In this paper, we examine **what factors related to individuals' *identities, experiences, and knowledge* influence their likelihood of rating harmful algorithmic biases**. We choose Image Search as our study context since (1) it is an everyday AI application and likely to be familiar to our participants; and (2) many past cases of successful user-driven algorithm audits have been in the context of image search (Guarino 2016; Kay, Matuszek, and Munson 2015; Lam et al. 2018). We set out to systematically investigate the effects of three sets of factors (demographics, everyday discrimination experiences, and social and technical knowledge) on users' likelihood of harmful bias ratings.

To answer these questions, we conducted **an online experiment on Prolific with 2,179 participants**, presenting three known biased image search results in race, gender, and sexual orientation as well as one neutral result, randomized out of a total pool of eight cases. Participants were asked to rate the level of harm of each case and then answer open-ended questions about their reasons behind their rating. They were then asked to answer several questions related to their identities and experiences. We found that users belonging to marginalized genders or sexual orientations were statistically significantly more likely to rate the image search result more severely harmful than users who were not members of genders or sexual orientations. Surprisingly, we did *not* find a statistically significant relationship between belonging to marginalized races and being more likely to rate the biased image search results as more severely harmful. In addition, users who have experienced discrimination in their everyday lives more frequently are more likely to rate algorithmic bias in image search as more harmful. We also found that users' relations with marginalized groups, their awareness of societal biases, media exposure to algorithmic and societal bias also positively impact their likelihood of rating harmful biases, while tech-savviness has no significant effect. Our qualitative findings further revealed the complicated reasons behind bias rating, suggesting design implications for developing better means to support future user-driven auditing.

2 Related Work

2.1 Engaging Diverse End-Users in Algorithm Testing and Auditing

In recent years, there have been significant efforts in the development of tools and processes to support auditing and documentation of potential harmful behaviors in a wide range of algorithmic systems, including social media recommendation systems, hiring algorithms, computer vision applications, and more (Noble 2018; Asplund et al. 2020; Sweeney 2013; Prates, Avelar, and Lamb 2020; Buolamwini and Gebru 2018; Hannak et al. 2014; Eslami et al. 2019; Metaxa et al. 2021; Sandvig et al. 2014). Past literature has generally focused on algorithm auditing by small groups of experts such as industry practitioners, researchers, and activists (Metaxa et al. 2021; Deng et al. 2023; Sandvig et al. 2014). However, because experts often audit the algorithm outside the context of everyday use, expert-led audits often miss serious issues that everyday users of algorithmic systems are able to detect once a system is deployed (Holstein et al. 2019; Shen et al. 2021; Cramer et al. 2018; Selbst et al. 2019). For instance, experts might lack the relevant cultural knowledge and lived experience to recognize and know where to look for certain kinds of harmful behaviors (DeVos et al. 2022; Holstein et al. 2019; Shen et al. 2021; Deng et al. 2022). Experts may overlook specific problematic behaviors of algorithms in internal red-teaming that can only be surfaced in real-world scenarios (Raji et al. 2020; Deng et al. 2023).

Recognizing this, there has been a shift towards "everyday algorithm auditing" where users organically coming together to uncover and raise awareness about harmful behaviors in algorithmic systems they use day-to-day (Shen et al. 2021; Eslami et al. 2019; Li et al. 2023). AI and HCI researchers have begun to explore the design of tools and processes to engage everyday users in surfacing problematic algorithm behaviors (Deng et al. 2023; Metaxa et al. 2021; Attenberg, Ipeirotis, and Provost 2015; Cabrera et al. 2021; Kiela et al. 2021; Nushi, Kamar, and Horvitz 2018; Ochigame and Ye 2021; Lam et al. 2022). In parallel, several major technology companies have also started experimenting with approaches that engage users in testing and auditing their AI services for problematic behaviors (Warkentin and Woodward 2022; Pistilli 2022; Cattell, Chowdhury, and Carson 2023). Three prominent examples include Twitter's "algorithmic bias bounty" to engage users in identifying harmful biases in its image cropping algorithm (Chowdhury and Williams 2021), OpenAI's feedback contests and red teaming efforts to engage end-users in providing feedback on problematic outputs with ChatGPT (AI 2022), and AI Village's Generative AI red teaming exercise at the DEFCON hacker convention (Cattell, Chowdhury, and Carson 2023).

Despite the growing efforts of engaging end-users in algorithm auditing, there is still a significant knowledge gap regarding effective methods of identifying and recruiting the right group of user-auditors. Past literature offers *initial qualitative evidence* based on small samples, indicating that users can leverage lived experience to uncover potential harmful biases (DeVos et al. 2022). However, as (Deng et al. 2023) highlighted, the current literature still lacks comprehensive understanding on how to identify specific user subgroups for algorithm auditing, how to assemble a diverse and representative group of users, and how to *minimize the burden* on already marginalized groups.

This paper addresses this gap by presenting *large scale, quantitative* evidence about how various factors – demographics, discrimination experiences and social and technical knowledge – might influence users' perceptions of and reactions to harmful algorithmic biases. This knowledge is vital for the effective recruitment of user-auditors and for designing systems that involve users in the auditing process.

2.2 Impacts of User Identity and Experience on Perceptions of Algorithmic Behavior

Researchers in social science have long explored the way that one's identity and lived experience might affect the

way they perceive and react to societal biases and stereotypes: in the context of interpersonal interactions, for instance, Kessler et al. observed that individuals from racial minorities, particularly those who have experienced racial discrimination, are more inclined to recognize systemic bias (Kessler, Mickelson, and Williams 1999); similarly, studies have indicated that individuals from marginalized genders or sexual orientations are more apt to perceive biases against their groups, possibly due to experiences of discrimination (Eagly and Kite 1987; Cobb et al. 2019; Herek 2009); in media representation, beyond personal identity, exposure to media perpetuating biases, such as mass media content presenting traditional gender stereotypes, might also increase individual's sensitivity towards information that contains biases (Durkin 1985; Hurtz and Durkin 2004; Devine 1989).

Our work extends past social psychology theories on the relationship between users' identity/experiences and bias perceptions in the ***emerging context of algorithmic auditing***, exploring the relationship between users' identities and experiences and their likelihood of harmful bias rating when interacting with algorithms.

Emerging research within AI and HCI has offerred some initial qualitative evidence of the impact of user identity and experience on perceptions of *algorithmic* behaviors, or in the context of AI annotation tasks that might affect the downstream outputs (DeVos et al. 2022; Lam et al. 2022; Qadri et al. 2023; Dennler et al. 2023; Wang, Harper, and Zhu 2020; Wenzel et al. 2023). Our study is the first large-scale, quantitative investigation into the factors that affect users' likelihood of harmful bias rating, complemented by qualitative analysis to understand the reasons behind users' ratings.

3 Hypotheses

3.1 Hypothesis (HP) 1a - 1c : Effects of Demographic Features

The field of social psychology has long established that individuals from marginalized demographics are more attuned to societal biases affecting their groups. For instance, research has shown that individuals from racial minority groups are more likely to identify systemic racial bias (Kessler, Mickelson, and Williams 1999; Cobb et al. 2019). Similarly, previous literature suggests women are more inclined to perceive gender discrimination (Eagly and Kite 1987), and individuals from marginalized sexual orientations are more prone to recognizing biases and stereotypes against their communities (Herek 2009). Also, previous research in AI and HCI advocated for avoiding overburdening the already marginalized groups (DeVos et al. 2022; Deng et al. 2023), we are motivated to evaluate the effect of a certain marginalized demographic on users' rating of harmful algorithmic bias not only against this specific demographic, but also other marginalized demographics. This led us to the following hypotheses: **HP 1a**. Users belonging to marginalized genders are more likely to detect harmful algorithmic bias in image search. **HP 1b**. Users belonging to marginalized sexual orientations are more likely to detect harmful algorithmic bias in image search. **HP 1c**. Users belonging to marginalized races are more likely to detect harmful algorithmic bias in image search.

3.2 Hypothesis (HP) 2: Effects of Everyday Discrimination Experience

Social psychologists have found that individuals are more likely to perceive societal bias and stereotypes against marginalized genders, sexual orientations, or races due to experiences of discrimination (Kessler, Mickelson, and Williams 1999; Eagly and Kite 1987; Cobb et al. 2019; Herek 2009). In the context of algorithm auditing, DeVos et al. (DeVos et al. 2022) found prior exposure to and experiences with harmful demographic bias have a large impact on types of bias people are able to identify. Using the Everyday Discrimination Scale, which measures the frequency and impact of perceived discrimination in daily life (Williams et al. 1997; Carter et al. 2017; Harnois et al. 2019), we hypothesized the following: **HP 2**. Users who have experienced discrimination in their everyday lives more frequently are more likely to detect harmful algorithmic bias and discrimination in image search.

3.3 Hypothesis (HP) 3a - 3e: Effects of Social and Technical Knowledge

Previous literature in social psychology found that media exposure to societal biases has a profound impact on how individuals perceive bias in information (Durkin 1985; Durkin and Hurtz 2004; Devine 1989). In algorithm auditing, DeVos et al. (DeVos et al. 2022) found that users' performance is also influenced by second-hand knowledge from close social relations and secondary sources of information., e.g. anecdotal accounts from friends and family, as well as media reports about societal and algorithmic biases. This led us to the following hypotheses: **HP 3a**. Users with relationships to marginalized demographics are more likely to detect harmful algorithmic bias in image search. **HP 3b**. Users with awareness of societal bias are more likely to detect harmful algorithmic bias in image search. **HP 3c**. Users with media exposure to societal bias are more likely to detect harmful algorithmic bias in image search. **HP 3d**. Users with media exposure to algorithmic bias are more likely to detect harmful algorithmic bias in image search. **HP 3e**. Users with perceived familiarity with algorithmic systems are more likely to detect harmful algorithmic bias in image search.

4 Method

We conducted a randomized survey experiment on Prolific. 2,221 participants completed the survey, reduced to 2,179 after cleaning. The study protocol was approved by our Institutional Review Board (IRB).

4.1 Experiment Design

We designed an online experiment where participants were shown four cases of Image Search results in random order, including three cases that previous literature has identified as biased against marginalized gender, racial, or sexual orientation groups, as well as one neutral case. We presented four cases for each participant by randomly selecting one from each set of two in an eight-case pool. We randomized

the case order to avoid any order effects, following prior work (Seltman 2013; Gorvine et al. 2017). Given that our hypotheses were preregistered to investigate the effects of identities and experience on user harmful bias ratings, we do not assign participants to treatment and control groups. Our *independent variables* were participants' demographics, everyday discrimination experience, and social and technical knowledge; our *dependent variable* was participants' ratings of how harmful they perceived the biases were in each case.

Case Selection: We chose six biased cases of image search results from previous literature. These cases have drawn wide attention due to one type of bias based on race, gender, or sexual orientation, as there have been rich discussions of harmful algorithmic biases in these dimensions. The selection process started with a small set of high-profile cases reported by previous literature. We iteratively reviewed these known cases and searched for new ones to reach six bias cases, using criteria such as they should be able to be presented via screenshots and should be easy to understood by participants. It is worth noting that some of these cases might be perceived as containing multiple types of biases, as the identification of harmful algorithmic biases often involves complex and subjective value judgments, and users may perceive those biased cases through multiple dimensions, as suggested by (DeVos et al. 2022). For the purpose of this study, we categorized these cases based on the primary bias identified in the literature, which served as our ground truth, into three *Bias Case sets* (see Table 1). Specifically, we included search results of "tree" and "flower" as our Neutral Case, to reduce any learning effects (Charness, Gneezy, and Kuhn 2012). The use of non-human subjects as neutral cases aligns with research designs used in previous studies (Schoth and Liossi 2017), with a goal to minimize the possibility of incorporating demographic biases that, while not broadly acknowledged or discussed, could still be raised by some study participants (DeVos et al. 2022). Following (DeVos et al. 2022), we presented the collection of images in the format of one snapshot. The complete survey for our experiment, the study procedure, and figures of all the cases can be found in Supplementary Material[1].

4.2 Operationalization

Dependent variable:

- **Harmfulness of Bias (7-point Likert Scale):** The dependent variable was measured using a 7-point Likert scale, where 1 = Totally unharmful bias, and 7 = totally harmful bias. We converted it to an ordered factor.

Independent variables:

- **Belonging to a marginalized gender, sexual orientation, or race (Binary Variables):** We created three variables to represent whether a participant belonged to a marginalized gender, sexual orientation, or race. The "Belongs to a non-marginalized demographic" level was the reference level in our ordinal regression model.
- **Chronicity of yearly everyday discrimination experienced (Binned Factor):** We operationalized responses to the short-version of the Everyday Discrimination Scale (Williams et al. 1997; Carter et al. 2017; Harnois et al. 2019) by first multiplying each item response by a number representing the frequency at which the participant said they had experienced a particular type of discrimination throughout the year, then summed the values across the different items, and binned the summed values into "low", "moderate", and "high" occurrence of yearly discrimination, following the "chronocity" method in previous literature (Michaels et al. 2019). "Low" of yearly discrimination was the reference level.
- **Relation to marginalized genders, sexual orientations, and races (Binary Variables):** We transformed participant responses of if they had family or friends who belonged to marginalized genders, sexual orientations, and races into binary variables, where 1 = yes, and 0 = they did not. "0" was the reference level.
- **Awareness of Societal Bias (Categorical, 6-Level Factor):** We transformed participant responses of their awareness of societal biases into a categorical factor variable with 6 levels, ranging from "Not aware at all" to "Very aware". "Not aware at all" was the reference level.
- **Media exposure to social bias information (Categorical, 6-Level Factor):** We transformed participant responses of their exposure to media about societal biases into a categorical factor variable with 6 levels from "Never" to "Daily". "Never" was the reference level.
- **Media exposure to algorithmic bias information (Categorical, 6-Level Factor):** We transformed participant responses of their exposure to media about algorithmic biases into a categorical factor variable with 6 levels from "Never" to "Daily". "Never" was the reference level.
- **Perceived familiarity with algorithmic systems (Categorical, 5-Level Factor):** We transformed participant responses of their perceived familiarity with algorithmic systems into a categorical factor variable with 5 levels, ranging from "Not familiar at all" to "Extremely familiar". "Not familiar at all" was the reference level.

4.3 Data Collection

Sampling Strategy: We conducted power analysis for each demographic group using US census data as a reference to determine sample sizes. For each demographic feature assessed - gender, race, and sexual orientation - we determined the estimated sample size, aiming for a 95% confidence level with a 5% margin of error. These calculations were based on each granular demographic classification within the categories of gender, race, and sexual orientation, and were aligned with the US census data's demographic ratios. As the categories provided by the Prolific prescreener did not align seamlessly with those of US census data, we collated the estimated sample sizes of conventionally privileged groups and lumped the estimated sample sizes for marginalized demographics into broader categories: male (N = 385), non-male (N = 421), white (N = 375), non-white (N = 567), heterosexual (N = 158), and non-heterosexual (N = 176).

[1]See our supplementary material at https://github.com/jiayin3zh/crowdauditbaseline-public.git

Case Sets	Cases of Image Search Results	Descriptions
Bias Case Set 1	"professor style" on Google Images	The results only showed men (Noble 2018).
	"doctor" on Google Images	The results only showed white men (Noble 2018).
Bias Case Set 2	"weddings" on Google Images	The results only showed heterosexual relationships (DeVos et al. 2022).
	"romantic couples" on Google Images	The results only showed heterosexual relationships (DeVos et al. 2022).
Bias Case Set 3	"babies" on Bing Images	The results only displayed white babies (Kleinman 2017).
	"CEO" on Google Images	The results showed all white men (Brekke 2015).
Neutral Case Set	"flower" on Google Images	The results are from "flower" on Google Image.
	"tree" on Google Images	The results are from "tree" on Google Image.

Table 1: Cases we used in our experiment. Prior literature has reported these cases as biased against marginalized genders, sexual orientations, and races. We presented four cases randomly, randomly selecting one case from each case set.

Participants: Using the Prolific pre-screener for filtration and anticipating some low-quality responses, we concurrently released six studies and recruited 5% more participants for each sample: male (N = 411), non-male (N = 448), white (N = 399), non-white (N = 600), heterosexual (N = 170), and non-heterosexual (N = 193). 2,221 participants were recruited to participate in a study termed "Asking for your opinion about algorithmic systems" on Prolific. We only allowed workers with a minimum approval rate of 90% to participate. Each participant was at least 18 years old, located in the US, and could participate only once. Participants were rewarded based on a $12 hourly rate and the average duration was 8 minutes and 35 seconds. Responses were rejected if their duration were 3 standard deviations below the mean, if they failed the attention check, or if the participant was asked to write a sentence but only put in several words. We excluded 7 participants who failed the attention check question (Q10), and 13 participants who reported that they had previously seen any of the cases. We used a Large Language Model (LLM) to filter low-quality, self-contradictory responses which were excluded from the final dataset. Our LLM application followed best validation practices: the model had a 98.7% accuracy and 0.974 kappa value based on our human-labeled subset. This led to an exclusion of 22 participants. Our final dataset contains 2,179 participants after the cleaning. We detail our use of the LLM and the table of participants' demographics in Supplementary Material[1].

4.4 Qualitative Analysis

To understand the rationale behind participants' ratings, we performed content analysis (Braun and Clarke 2006) on the open-ended questions, following (Emami-Naeini et al. 2021). Due to the large size of our data, we randomly sampled a subset of 324 responses, following Engel et al. (2022). For the three biased cases, we categorized participants' responses as harmful, unharmful, or neutral, based on their harmfulness ratings. This resulted in nine subsets of responses (3 cases x 3 categories). We randomly selected 36 responses from each subset, which resulted in a total of 324 responses (5.1% of all). Two researchers read through the dataset individually, held weekly discussions, and developed a codebook. They inductively coded a random sample of 36 responses and reached a Cohen's Kappa value of 0.85, above the recommended 0.70 (McDonald, Schoenebeck, and Forte 2019). The remaining 288 responses was coded by one.

5 How Various Factors Influence Users' Rating of Harmful Algorithmic Bias

We detail how different factors impacted how users rated the severity of harmful algorithmic bias in each case. We used Cumulative Link Models (CLM) - a type of ordinal regression for examining ordered categorical data - to test our hypotheses. We constructed three Models a, b and c, each for each case set to test HP 1a - 1c, 2, and 3a - 3e (see Table 2).

5.1 Effects of Demographics

Marginalized Genders (HP 1a): Of those participants whose genders are marginalized in society, 69.1%, 41.6%, and 74.5% respectively rated the image search results in Case #1, #2, and #3 as either totally, very, or somewhat harmful. Of those participants whose genders are not marginalized in society, 45.2%, 23.5%, and 56.0% respectively rated the image search results in Case #1, #2, and #3 as either totally, very, or somewhat harmful. We found a statistically significant relationship between **marginalized genders** and participants rating Case #1, #2, or #3 as harmfully biased. The estimated coefficient for **marginalized genders** was 0.653 (95% CI: [0.494, 0.812], $p<0.001$) in case #1; 0.443 (95% CI: [0.283, 0.602], $p<0.001$) in case #2; and, 0.601 (95% CI: [0.441, 0.760], $p<0.001$) in Case #3. For people who belong to marginalized genders, the odds (based on exponentiation of the significant coefficients, following (Grano et al. 2020)) of being more likely to rate the case as harmful is 1.89 (Case #1), 1.56 (Case #2), or 1.82 (Case #3) times that of people not belonging to marginalized genders. **Thus, H1a is supported.**

Marginalized Sexual Orientations (HP 1b): Of those participants whose sexual orientations are marginalized in society, 67.1%, 47.2%, and 75.9% respectively rated the image search results in Case #1, #2, and #3 as either totally, very, or somewhat harmful. Of those participants whose sexual orientations are not marginalized in society, 56.5%, 27.7%, and 61.6% respectively rated the image search results in Case #1, #2, and #3 as either totally, very, or somewhat harmful. We found a statistically significant

	Model a *(Case Set 1)*	**Model b** *(Case Set 2)*	**Model c** *(Case Set 3)*
Demographic Group Status:			
Marginalized Gender	0.653*** [0.494, 0.812]	0.443*** [0.283, 0.602]	0.601*** [0.441, 0.760]
Marginalized Sexual Orientation	0.224* [0.028, 0.420]	0.610*** [0.410, 0.809]	0.432*** [0.236, 0.628]
Marginalized Race	-0.107 [-0.276, 0.062]	-0.220* [-0.392, -0.049]	-0.021 [-0.190, 0.148]
Relationships to Marginalized Demographics:			
Relationships to Gender Marginalized	0.243* [0.035, 0.451]	0.282** [0.069, 0.496]	0.338** [0.128, 0.547]
Relationships to Sexual Orientation Marginalized	0.333** [0.128, 0.538]	0.329** [0.119, 0.538]	0.245* [0.038, 0.453]
Relationships to Race Marginalized	-0.030 [-0.259, 0.199]	-0.144 [-0.377, 0.088]	0.154 [-0.075, 0.383]
Perceived Familiarity to Algorithmic System:			
Extremely familiar	-0.306 [-0.783, 0.170]	-0.330 [-0.826, 0.166]	-0.399 [-0.884, 0.086]
Moderately familiar	-0.192 [-0.631, 0.246]	-0.226 [-0.681, 0.229]	-0.294 [-0.740, 0.152]
Moderately not familiar	-0.296 [-0.781, 0.189]	-0.093 [-0.597, 0.412]	-0.312 [-0.804, 0.179]
Neither familiar nor not familiar	-0.205 [-0.671, 0.262]	-0.144 [-0.628, 0.341]	-0.284 [-0.760, 0.193]
Awareness of Societal Biases:			
Very aware	0.599 [-0.312, 1.511]	1.189* [0.155, 2.222]	1.619** [0.626, 2.612]
Somewhat aware	0.682 [-0.176, 1.540]	1.059* [0.072, 2.047]	1.578*** [0.641, 2.515]
Neither aware or not aware	1.087* [0.256, 1.918]	1.300** [0.337, 2.262]	2.080*** [1.168, 2.992]
Not very aware	1.395** [0.557, 2.233]	1.275** [0.308, 2.243]	2.428*** [1.510, 3.347]
Media Exposure to Societal Bias:			
Daily	0.558+ [-0.048, 1.164]	0.351 [-0.274, 0.977]	0.441 [-0.160, 1.042]
Weekly	0.662* [0.070, 1.253]	0.454 [-0.157, 1.065]	0.632* [0.046, 1.218]
Monthly	0.462 [-0.143, 1.068]	0.335 [-0.287, 0.958]	0.726* [0.127, 1.325]
A few times a year	0.304 [-0.307, 0.914]	0.258 [-0.374, 0.890]	0.389 [-0.216, 0.994]
I don't know	0.383 [-0.305, 1.070]	0.467 [-0.243, 1.177]	0.350 [-0.336, 1.037]
Media Exposure to Algorithmic Bias:			
Daily	0.053 [-0.306, 0.412]	0.102 [-0.263, 0.466]	0.228 [-0.130, 0.586]
Weekly	-0.057 [-0.351, 0.237]	0.002 [-0.299, 0.303]	0.071 [-0.223, 0.365]
Monthly	0.115 [-0.170, 0.399]	0.227 [-0.063, 0.517]	0.198 [-0.085, 0.480]
A few times a year	0.320* [0.057, 0.583]	0.032 [-0.237, 0.302]	0.299* [0.038, 0.561]
I don't know	0.022 [-0.263, 0.308]	0.062 [-0.229, 0.354]	0.290* [0.004, 0.576]
Yearly Discrimination Chronicity:			
Yearly Discrimination	0.231*** [0.097, 0.366]	0.389*** [0.208, 0.409]	0.245*** [0.087, 0.346]
Num.Obs.	2179	2179	2179
AIC	7586.9	7181.0	7461.7
BIC	7854.2	7442.6	7723.3
RMSE	4.44	3.42	4.68
Note:			+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Models predicting more severe harmfulness rating from demographics, everyday discrimination experience, and social and technical knowledge. Bracket next to each coefficient is the Confidence Interval: [Confidence Interval 95%: Low, High].

relationship between belonging to **marginalized sexual orientations** and participants rating the image search results in Case #1, #2, and #3 respectively as harmfully biased. For brevity, we refer readers to Table 2 for the coefficients for each case. For people who belong to marginalized sexual orientations, the odds of being more likely to rate the case as harmful is 1.25 (Case #1), 1.84 (Case #2), or 1.54 (Case #3) times that of people not belonging to marginalized sexual orientations. **Thus, H1b is supported.**

Marginalized Races (HP 1c): Of those participants whose races are marginalized in society, 54.9%, 27.6%, and 62.9% respectively rated the image search results in Case #1, #2, and #3 as either totally, very, or somewhat harmful. Of those participants whose races are not marginalized in society, 61.6%, 35.2%, and 66.2% respectively rated the image search results in Case #1, #2, and #3 as either totally, very, or somewhat harmful. We did not find a statistically significant relationship between belonging to **marginalized races** and participants rating the image search results in Case #1 and #3 as harmfully biased. We did find a statistically significant relationship in Case #2. However, the sign of the coefficient for Model b for marginalized race was negative in Case #2, meaning participants belonging to a marginalized race demographic were less likely to rate the image results as more harmfully biased. The estimated coefficient for **marginalized races** was -0.220 (95% CI: [-0.392, -0.049], $p<0.05$) in Case #2. **Thus, H1c is not supported.**

5.2 Effects of Everyday Discrimination Experience

We found a statistically significant relationship between *"Yearly Discrimination Chronocity"* **(experiencing *more* everyday discrimination throughout the year)** and participants rating the image search results in Cases #1, #2, and #3 respectively as harmfully biased. The estimated

coefficients were 0.231 (95% CI: [0.097, 0.366], p<0.001) in Case #1; 0.389 (95% CI: [0.208, 0.409], p<0.001) in Case #2; and, 0.245 (95% CI: [0.087, 0.346], p<0.001) in Case #3. Readers should interpret the positive coefficient to mean that respondents tend to give a higher harmfulness rating to the image search results as the perceived yearly discrimination increases. **Thus, Hypothesis 2 is supported.**

5.3 Effects of Social and Technical Knowledge

Relationships to Marginalized Demographics (HP 3a): In Case #1, #2, or #3, we found a statistically significant relationship between participants **having relationships to marginalized genders** and participants rating the image search results as harmfully biased. In Case #1, #2, or #3, we also found a statistically significant relationship between participants **having relationships to marginalized sexual orientations** and participants rating the case as harmful. For brevity, we refer readers to Table 2 for the coefficients for each case. In Case #1, #2, or #3, we did not find a statistically significant relationship between participants **having relationships to marginalized races** and harmfulness rating. **Thus, H3a is partially supported by having relationships to marginalized genders and having relationships to marginalized sexual orientations, and *not* supported by having relationships to marginalized races.**

Awareness of Societal Bias (HP 3b): In Case #1, we found a statistically significant relationship only between participants reporting they were *"Neither aware or not aware"* and *"Not very aware"* in **awareness of societal biases** and their rating the image search results as harmfully biased. In Case #2 and #3, we found a statistically significant relationship between ***every level*** of participants' **awareness of societal biases** and their rating the image search results as harmfully biased. For brevity, we refer readers to Table 2 for the coefficients for each level of awareness for each case. **Thus, H3b is partially supported in Case #1, and supported in Case #2 and #3.**

Media Exposure to Societal Bias (HP 3c): In Case #1, we found a statistically significant relationship only between *"weekly"* **media exposure to societal bias** and participants rating the image search results as harmfully biased. The estimated coefficient for *"weekly"* was 0.662 (95% CI: [0.070, 1.253], p<0.05) in Case #1. In Case #3, we found a statistically significant relationship only between *"weekly"* or *"monthly"* **media exposure to societal bias** and participants rating the image search results as harmfully biased. The estimated coefficient for *"weekly"* was 0.632 (95% CI: [0.046, 1.218], p<0.05) and the estimated coefficient for *"monthly"* was 0.726 (95% CI: [0.127, 1.325], p<0.05) in Case #3. **Thus, H3c is partially supported in Case #1 and #3.**

Media Exposure to Algorithmic Bias (HP 3d): In Case #1 and #3, we found a statistically significant relationship between *"a few times a year"* **media exposure to algorithmic bias** and participants rating the image search results as harmfully biased. In Case #1, the estimated coefficient for *"a few times a year"* was 0.320 (95% CI: [0.057, 0.583], p<0.05). In Case #3, the estimated coefficient for *"a few times a year"* was 0.299 (95% CI: [0.038, 0.561], p<0.05). **Thus, H3d is partially supported in Case #1 and #3.**

Perceived Familiarity with Algorithmic Systems (HP 3e): We find no statistically significant relationship. **Thus, H3e is not supported.**

6 Why Users See Biases the Way They Do

We categorized the 324 responses into three major themes: (1) unharmful; (2) harmful; and (3) neutral. Each comprised 108 responses. Our analysis revealed a multifaceted interplay of factors such as societal norms, perceptions of diversity, and the depth of social and technical understanding.

6.1 Perceived as Harmful

Our findings highlight three primary reasons participants deemed an algorithmic outcome harmful: 1) **Devaluation of Marginalized Groups.** Some (19/108) argued that the image search results can harm the affected minority groups by devaluing them. The results were perceived as fostering feelings of inferiority, exclusion, or discrimination among marginalized groups: *"People of color may feel their babies are less valued in society than white babies. This causes psychological harm"* (P80). 2) **Erasure of Marginalized Groups.** Some highlighted that by only showing white, male, heterosexual groups, these biased image search results actually ignore and/or erase the existence of marginalized groups (46/108). One explained, *"there are lots of black and female CEO's not pictured (here)"* (P115). 3) **Generating Allocative Harms.** Some (18/108) stressed that the image search results could generate direct negative societal impact for marginalized groups by eroding their confidence and making them feel less capable of achieving their goals in society, connecting "representational harms" with "allocative harms" (Crawford 2017): *"This might enforce the idea that women can't be doctors, which would harm girls who aspire to be doctors"* (P26).

6.2 Perceived as Unharmful

Participants who downplayed harmfulness of biased image search results generally held three perspectives: 1) **Reflection of Social Reality.** Although some noted that marginalized groups were less represented, they (10/108) believed the search results were merely echoing existing societal structures, and therefore, were not harmful. One felt the "doctor" case simply mirrored the real-world demographic makeup (P119). 2) **Perceived Adequate Diversity.** Some (31/108) believed that these results were already adequately diverse, particularly in regard to racial representation. In many cases, their assessments often centered on racial diversity, while overlooking potential biases against other marginalized groups. One felt, *"I believe the bias and discrimination this algorithmic system is generating is somewhat unharmful because it is inclusive of other races and nationalities"* (P143). Even when acknowledging the lack of diversity in other social dimensions (e.g., sexual orientation), some still deemed the overall situation as unharmful due to the presence of racial diversity: *"I believe it is not*

harmful because I see a good amount of different weddings of different races, although there are more straight couples" (P154). 3) **Technical Interpretation and Social Disconnect.** Some (47/108) evaluated search algorithms from a purely technical standpoint, detaching social implications. As long as the search results were relevant to the query, the algorithm was functioning appropriately (P201). They felt that the algorithm was performing as expected and shifted the responsibility for guiding search outcomes to humans: *"It is an innocent, AI-driven search. If you want to see certain results, you need to change the term"* (P186).

6.3 Neutral Perception

Some participants could not decide whether the image search results contained harmful bias. 1) **Ambiguity over Adequate Diversity.** Some (14/108) acknowledged the lack of representation but were not sure how much diversity is enough: *"It becomes an issue when trying to describe what is enough that makes it potentially harmful"* (P301). 2) **Disconnection between Diversity and Impact.** Some (36/108) admitted that although they noticed issues with limited representations of minorities, they were not sure whether such results are generating harmful impacts.

7 Discussion

Broadening Participation for User-Engaged Algorithm Auditing. Our findings provide practical insights for addressing the challenges researchers and practitioner faced when implementing user-engaged algorithm auditing, such as identifying relevant user subgroups, recruiting appropriate end-users, and assigning tasks to gain actionable insights (Deng et al. 2023; Raji et al. 2020), as well as avoiding overburdening marginalized groups due to the offensive nature of many AI biases(Steiger et al. 2021). First, we provide quantitative evidence confirming that individuals from marginalized groups, particularly gender and sexual orientation minorities, are more likely to rate algorithmic biases as more harmful in image search. This implies that practitioners should prioritize recognizing and amplifying the voices of marginalized groups in their processes. Our results also suggest potential "allies" whom practitioners can recruit to assist minorities in algorithm auditing. Specifically, our findings indicate that individuals who have experienced everyday discrimination, those who possess an understanding of societal bias, as well as those who have been exposed to algorithmic and societal bias through the media, are more likely to rate algorithmic biases as more harmful. To potentially relieve the burdens from marginalized groups, practitioners should consider recruiting additional participants from these social groups to support improving AI products and services for minorities. Expanding the scope to identify more potential allies widens the pool of participants, streamlining the recruitment process without compromising the quality of the audit (Raji and Buolamwini 2019; Costanza-Chock, Raji, and Buolamwini 2022).

The Complexity in Algorithmic Bias Sensitivity. Our work has broadened our understanding of how people's demographics and experiences affect their likelihood to rate systematic or social biases, extending past literature in social psychology to the emerging domain of algorithm auditing. Within this new algorithmic context, our quantitative results indicate that marginalized genders or sexual orientations were more likely to rate algorithmic bias as more harmful than other non-marginalized participants. However, there is no statistically significant relationship between belonging to marginalized races and being more likely to rate the biased image search results as more severely harmful. We are unsure of the underlying reasons for this disparity and believe they warrant further investigation. Here, we discuss a potential reason, based on evidence from our qualitative analysis. Users sometimes downplayed biases related to gender and sexual orientation if they perceived there was sufficient racial representation. This, coupled with the recent growing awareness of social bias and increasing media coverage on discrimination, may have influenced users to be more sensitive to bias, regardless of their own races. Our quantitative data also revealed that having heightened awareness of societal biases increases the likelihood of rating biased cases as harmful. It is worth noting that this observation is speculative and requires further research for confirmation.

Raising Awareness and Training the Next Generation of User-Auditors. One way practitioners can support users in algorithm audits is by drawing upon the insights from other users to offer *community guidance*. While our qualitative findings reveal a complex picture of the rationale behind participants' ratings, our participants have already offered important insights. Some drew from their lived experiences to illustrate the harms encoded in these cases: *"As a female doctor myself, this is harmful to the young girl's view and makes people feel they cannot become doctors because they believe every one of them is male"*(P12). We envision that practitioners can gather these community insights and share them with user groups, to promote awareness and foster critical AI literacy (Strauß 2021; Long and Magerko 2020). In addition, we also envision that practitioners can better support users in harmful bias rating by providing more *structural guidance*. For instance, in our experiment, some participants rated image search results that contain biases against women as harmless because they focused solely on evaluating racial representation while neglecting other forms of bias and discrimination (P143). To address this issue, practitioners could consider offering a comprehensive taxonomy of harmful biases or checklists, as (Shelby et al. 2023), to scaffold the process by assisting users in conducting more thorough assessments of various types of bias.

8 Conclusion

In this paper, we investigated the factors related to individuals' identities, experiences, and social and technical knowledge that influence their likelihood of rating harmful algorithmic biases. Our results identified potential allies we can enlist to support user-engaged algorithm auditing, thereby reducing the burden on marginalized social groups. We also offer insights into the complex reasons behind harmful bias rating. Our paper provides practical guidelines for designing future methods to better support users in algorithm auditing.

Acknowledgements

We thank our participants for their time and input that shaped this research. We thank Geoff Kaufman, Andy Lee, and anonymous reviewers for their insightful feedback on the study design and paper draft. This work was supported by the National Science Foundation (NSF) program on Fairness in AI in collaboration with Amazon under Award No. IIS-2040942, an award from Google Research and a fellowship from Microsoft Research. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of NSF or other funding agencies.

References

AI, O. 2022. ChatGPT Feedback Contest: Official Rules.

Albergotti, R. 2020. Black creators sue YouTube, alleging racial discrimination. *The Washington Post*.

Asplund, J.; Eslami, M.; Sundaram, H.; Sandvig, C.; and Karahalios, K. 2020. Auditing race and gender discrimination in online housing markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 24–35.

Attenberg, J.; Ipeirotis, P.; and Provost, F. 2015. Beat the machine: Challenging humans to find a predictive model's "unknown unknowns". *Journal of Data and Information Quality (JDIQ)*, 6(1): 1–17.

Bensinger, G.; and Albergotti, R. 2019. YouTube discriminates against LGBT content by unfairly culling it, suit alleges. *The Washington Post*.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Brekke, K. 2015. Google Image Search Has a Gender Bias Problem. 1. Accessed: 2023-09-05.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Cabrera, Á. A.; Druck, A. J.; Hong, J. I.; and Perer, A. 2021. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–22.

Carter, R. T.; Forsyth, J. M.; Mazzula, S. L.; and Williams, B. 2017. The prevalence of discrimination across racial groups in contemporary America: Results from a nationally representative sample of adults. *PloS one*, 12(8): e0183356.

Cattell, S.; Chowdhury, R.; and Carson, A. 2023. AI Village at DEF CON announces largest-ever public Generative AI Red Team.

Charness, G.; Gneezy, U.; and Kuhn, M. A. 2012. Experimental methods: between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81: 1–8.

Chowdhury, R.; and Williams, J. 2021. Introducing Twitter's first algorithmic bias bounty challenge. *URl: https://blog. twitter. com/engineering/en_us/topics/insights/2021/algorithmic-bias-bountychallenge*.

Cobb, C. L.; Meca, A.; Branscombe, N. R.; Schwartz, S. J.; Xie, D.; Zea, M. C.; Fernandez, C. A.; and Sanders, G. L. 2019. Perceived discrimination and well-being among unauthorized Hispanic immigrants: The moderating role of ethnic/racial group identity centrality. *Cultural Diversity and Ethnic Minority Psychology*, 25(2): 280.

Costanza-Chock, S.; Raji, I. D.; and Buolamwini, J. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583.

Cramer, H.; Garcia-Gathright, J.; Springer, A.; and Reddy, S. 2018. Assessing and addressing algorithmic bias in practice. *Interactions*, 25(6): 58–63.

Crawford, K. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Deng, W. H.; Guo, B.; Devrio, A.; Shen, H.; Eslami, M.; and Holstein, K. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.

Deng, W. H.; Nagireddy, M.; Lee, M. S. A.; Singh, J.; Wu, Z. S.; Holstein, K.; and Zhu, H. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 473–484.

Dennler, N.; Ovalle, A.; Singh, A.; Soldaini, L.; Subramonian, A.; Tu, H.; Agnew, W.; Ghosh, A.; Yee, K.; Peradejordi, I. F.; et al. 2023. Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 375–386.

Devine, P. G. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, 56(1): 5.

DeVos, A.; Dhabalia, A.; Shen, H.; Holstein, K.; and Eslami, M. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. *CHI Conference on Human Factors in Computing Systems*.

Durkin, K. 1985. *Television, Sex Roles and Children: A Developmental Social Psychological Account.* ERIC.

Durkin, K.; and Hurtz, K. 2004. Media exposure and perceptions of bias and discrimination. *Journal of Media Psychology*.

Eagly, A. H.; and Kite, M. E. 1987. Are stereotypes of nationalities applied to both women and men? *Journal of personality and social psychology*, 53(3): 451.

Emami-Naeini, P.; Dheenadhayalan, J.; Agarwal, Y.; and Cranor, L. F. 2021. Which Privacy and Security Attributes Most Impact Consumers' Risk Perception and Willingness to Purchase IoT Devices? In *2021 IEEE Symposium on Security and Privacy (SP)*, 519–536.

Engel, K.; Hua, Y.; Zeng, T.; and Naaman, M. 2022. Characterizing Reddit Participation of Users Who Engage in the QAnon Conspiracy Theories. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–22.

Eslami, M.; Vaccaro, K.; Lee, M. K.; Elazari Bar On, A.; Gilbert, E.; and Karahalios, K. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Eubanks, V. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Gorvine, B. J.; Rosengren, K. S.; Stein, L.; and Biolsi, K. 2017. *Research methods: From theory to practice*. Oxford University Press.

Grano, G.; De Iaco, C.; Palomba, F.; and Gall, H. C. 2020. Pizza versus Pinsa: On the Perception and Measurability of Unit Test Code Quality. In *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 336–347.

Guarino, B. 2016. Google faulted for racial bias in image search results for black teenagers. *Washington Post*, 6: 2016.

Hannak, A.; Soeller, G.; Lazer, D.; Mislove, A.; and Wilson, C. 2014. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, 305–318.

Harnois, C. E.; Bastos, J. L.; Campbell, M. E.; and Keith, V. M. 2019. Measuring perceived mistreatment across diverse social groups: An evaluation of the Everyday Discrimination Scale. *Social Science & Medicine*, 232: 298–306.

Herek, G. M. 2009. Sexual stigma and sexual prejudice in the United States: A conceptual framework. In *Contemporary perspectives on lesbian, gay, and bisexual identities*, 65–111. Springer.

Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudik, M.; and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16.

Hurtz, W.; and Durkin, K. 2004. The Effects of Gender-Stereotyped Radio Commercials 1. *Journal of applied social psychology*, 34(9): 1974–1992.

Jiang, J. A.; Scheuerman, M. K.; Fiesler, C.; and Brubaker, J. R. 2021. Understanding international perceptions of the severity of harmful content online. *PloS one*, 16(8): e0256762.

Kay, M.; Matuszek, C.; and Munson, S. A. 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*, 3819–3828.

Kessler, R. C.; Mickelson, K. D.; and Williams, D. R. 1999. The prevalence, distribution, and mental health correlates of perceived discrimination in the United States. *Journal of health and social behavior*, 208–230.

Kiela, D.; Bartolo, M.; Nie, Y.; Kaushik, D.; Geiger, A.; Wu, Z.; Vidgen, B.; Prasad, G.; Singh, A.; Ringshia, P.; et al. 2021. Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337*.

Kingsley, S.; Sinha, P.; Wang, C.; Eslami, M.; and Hong, J. I. 2022. " Give Everybody [..] a Little Bit More Equity": Content Creator Perspectives and Responses to the Algorithmic Demonetization of Content Associated with Disadvantaged Groups. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–37.

Kleinman, Z. 2017. Artificial intelligence: How to avoid racist algorithms.

Kumar, D.; Kelley, P. G.; Consolvo, S.; Mason, J.; Bursztein, E.; Durumeric, Z.; Thomas, K.; and Bailey, M. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, 299–318.

Lam, M. S.; Gordon, M. L.; Metaxa, D.; Hancock, J. T.; Landay, J. A.; and Bernstein, M. S. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Lam, O.; Broderick, B.; Wojcik, S.; and Hughes, A. 2018. Gender and jobs in online image searches.

Li, R.; Kingsley, S.; Fan, C.; Sinha, P.; Wai, N.; Lee, J.; Shen, H.; Eslami, M.; and Hong, J. 2023. Participation and Division of Labor in User-Driven Algorithm Audits: How Do Everyday Users Work together to Surface Algorithmic Harms? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.

Long, D.; and Magerko, B. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–16.

McDonald, N.; Schoenebeck, S.; and Forte, A. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.*, CSCW: 72:1–72:23.

Metaxa, D.; Park, J. S.; Robertson, R. E.; Karahalios, K.; Wilson, C.; Hancock, J.; Sandvig, C.; et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4): 272–344.

Michaels, E.; Thomas, M.; Reeves, A.; Price, M.; Hasson, R.; Chae, D.; and Allen, A. 2019. Coding the Everyday Discrimination Scale: implications for exposure assessment and associations with hypertension and depression among a cross section of mid-life African American women. *J Epidemiol Community Health*, 73(6): 577–584.

Noble, S. U. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Nushi, B.; Kamar, E.; and Horvitz, E. 2018. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, 126–135.

Ochigame, R.; and Ye, K. 2021. Search Atlas: Visualizing Divergent Search Results Across Geopolitical Borders. In *Designing Interactive Systems Conference 2021*, 1970–1983.

OpenAI. 2022. OpenAI: Our approach to alignment research.

Pistilli, G. 2022. HuggingFace announcedthe new feature to flag any Model, Dataset, or Space on the Hub.

Prates, M. O.; Avelar, P. H.; and Lamb, L. C. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10): 6363–6381.

Qadri, R.; Shelby, R.; Bennett, C. L.; and Denton, E. 2023. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 506–517.

Raji, I. D.; and Buolamwini, J. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678.

Schoth, D. E.; and Liossi, C. 2017. A systematic review of experimental paradigms for exploring biased interpretation of ambiguous information with emotional and neutral associations. *Frontiers in psychology*, 8: 171.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.

Seltman, H. 2013. Threats to your experiment. *Seltman, HJ, Experimental Design and Analysis*.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741.

Shen, H.; DeVos, A.; Eslami, M.; and Holstein, K. 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): 1–29.

Steiger, M.; Bharucha, T. J.; Venkatagiri, S.; Riedl, M. J.; and Lease, M. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–14.

Strauß, S. 2021. " Don't let me be misunderstood": Critical AI literacy for the constructive use of AI technology. *TATuP-Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis/Journal for Technology Assessment in Theory and Practice*, 30(3): 44–49.

Sweeney, L. 2013. Discrimination in online ad delivery. *Queue*, 11(3): 10–29.

Wang, R.; Harper, F. M.; and Zhu, H. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Warkentin, T.; and Woodward, J. 2022. AI Test Kitchen.

Wenzel, K.; Devireddy, N.; Davison, C.; and Kaufman, G. 2023. Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–14.

Williams, D. R.; Yu, Y.; Jackson, J. S.; and Anderson, N. B. 1997. Racial differences in physical and mental health: Socio-economic status, stress and discrimination. *Journal of Health Psychology*, 2(3): 335–351.