Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior

Alicia DeVos
Carnegie Mellon University
Pittsburgh, PA, USA
adevos@andrew.cmu.edu

Aditi Dhabalia
Carnegie Mellon University
Pittsburgh, PA, USA
aditidhabalia@gmail.com

Hong Shen
Carnegie Mellon University
Pittsburgh, PA, USA
hongs@andrew.cmu.edu

Kenneth Holstein*
Carnegie Mellon University
Pittsburgh, PA, USA
kjholste@andrew.cmu.edu

Motahhare Eslami*
Carnegie Mellon University
Pittsburgh, PA, USA
meslami@andrew.cmu.edu

ABSTRACT

Recent work in HCI suggests that users can be powerful in surfacing harmful algorithmic behaviors that formal auditing approaches fail to detect. However, it is not well understood how users are often able to be so effective, nor how we might support more effective user-driven auditing. To investigate, we conducted a series of think-aloud interviews, diary studies, and workshops, exploring how users find and make sense of harmful behaviors in algorithmic systems, both individually and collectively. Based on our findings, we present a process model capturing the dynamics of and influences on users' search and sensemaking behaviors. We find that 1) users' search strategies and interpretations are heavily guided by their personal experiences with and exposures to societal bias; and 2) collective sensemaking amongst multiple users is invaluable in user-driven algorithm audits. We offer directions for the design of future methods and tools that can better support user-driven auditing.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Computer supported cooperative work**; **Empirical studies in HCI**; • **Information systems** → *Web searching and information discovery*.

KEYWORDS

User-Driven Algorithm Auditing; Auditing Algorithms, Algorithmic Bias; Algorithmic Harm; Fair Machine Learning; Sensemaking; Search; Qualitative Methods; Collective Work

ACM Reference Format:
Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3491102.3517441

1 INTRODUCTION

The presence of biases and inequities in algorithmic systems has led researchers to develop new approaches for algorithm auditing to detect biased, discriminatory, or otherwise harmful behaviors[1] (e.g., [17, 22, 31, 46, 53, 59, 62, 66, 74, 82]). Typically these auditing techniques are led by experts such as researchers, activists, industry practitioners, and government agencies [20]. For example, in "scraping audit" techniques, experts query an algorithmic system and investigate and compare the outputs [66]. As another example, in a "sock puppet audit", experts use computer programs to impersonate different types of users, inject fake data into the system, and evaluate the results [66]. Although expert-led auditing approaches have been greatly impactful, the absence of the actual context of use and everyday users in the auditing process can still result in major blindspots in practice. As well as experts' cultural blindspots, social dynamics, changing norms, and new circumstances can hinder experts' detection of many types of algorithmic biases and harms [22, 35, 45, 69, 71, 81]. In contrast, recent years have seen many cases in which *users* uncover and raise awareness about harmful algorithmic behaviors that they encounter while interacting with algorithmic systems [31, 32, 44, 71]. Some examples include Twitter users detecting racial bias in Twitter's image cropping algorithm [44], Apple Card holders uncovering gender bias in Apple's credit limit algorithm [77], and a group of YouTubers reporting the demonetization of LGBTQ+ content by the YouTube recommendation algorithm [64].

Inspired by these observations, researchers are beginning to explore ways to harness the power of users and crowds in order to overcome limitations of expert-led algorithm auditing approaches [71]. For instance, researchers evaluated personalization impacts of search algorithms as well as price steering algorithms on e-commerce websites by crowdsourcing the outputs of these algorithms and evaluating the results [41, 42]. Another related line of research has developed prototype interfaces and crowdsourcing

*Both authors contributed equally to this research.

[1] When relevant, we differentiate between harmful biases and harmful behaviors more broadly in algorithmic systems. While these categories often overlap, not all algorithmic harms are best understood as "bias," and not all algorithmic biases are necessarily harmful [13].

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3517441

pipelines intended to support people in actively searching for algorithmic errors, biases, or harmful behaviors (e.g., [10, 18, 23, 60, 73]).

While these efforts have demonstrated promising results, it remains poorly understood *how* people go about searching for and making sense of harmful algorithmic behavior, let alone how they are often able to be so effective at doing so. In order to guide the design of more effective supports for user-driven algorithm auditing processes, it is critical to understand how these processes unfold, what search and sensemaking strategies users are likely to use, and where these strategies may be likely to break down. To begin to address these issues, in this paper, we explore the question: **How do people find, make sense of, and evaluate potentially harmful algorithmic behavior?**

We conducted a three-phase study with 23 everyday users of algorithmic systems to investigate how they surface and make sense of potentially harmful algorithmic behaviors. In the first phase of this study, we conducted think-aloud interview sessions in which individual participants were tasked with actively searching for and making sense of potentially harmful algorithmic behaviors in the context of Google Images search engine. To complement these live observations, participants then took part in a 14-day diary study in which they were asked to report potentially harmful algorithmic behaviors that they found both through active searching and incidentally during the course of their day-to-day interactions with algorithmic systems. Finally, we invited participants to participate in group workshop sessions and work collectively to interpret and discuss some of the cases that were uncovered during the diary study.

Throughout our study, participants showed great ability to detect and reason about potentially harmful algorithmic biases. For example, participants submitted a total of 160 evidence-supported reports of potential biases or other harmful algorithmic behaviors in the diary study, spanning gender, sexual orientation, race/ethnicity, socioeconomic class, age, body type, disability, religion, politics, and combinations of these. Our analyses yielded a three-stage model capturing the process by which people uncover biased and harmful algorithmic behaviors: 1) *search inspiration* encompasses ways in which participants came up with ideas for where to look for potential biases in an algorithmic system, 2) *sensemaking* involves the ways that participants understood and evaluated algorithmic behaviors that might be harmfully biased, and 3) *remediation* covers participant actions and desires to mitigate the ramifications of harmful algorithmic bias. Connected to each stage of this process, we uncovered two common influences on the process across participants: knowledge and beliefs, and platform affordances. Participants used their prior knowledge and beliefs related to biased and harmful algorithmic behavior (i.e., their conception of bias and harm, exposure and experience, expectations and values, and folk theories) to make sense of problematic algorithmic behaviors they encountered. Participants' abilities to detect and make sense of such behaviors were enabled and guided by platform affordances.

Based on our findings, we offer directions for the design of future methods and tools that can better support user-driven algorithm auditing. We discuss benefits, limitations, and challenges of user-driven algorithm auditing, informed by our process model. In doing so, we look to ways that future user-driven algorithm auditing platforms and tools might be designed to best leverage the benefits, mitigate the limitations, and navigate the tensions.

In summary, this work contributes the following to the HCI and CSCW communities:

- The first in-depth empirical investigation of how users find and make sense of harmful behaviors in algorithmic systems, both individually and collectively
- A conceptual process model capturing the ways that users find and make sense of biased and harmful algorithmic behaviors, developed through an analysis of our study participants' actions and explanations
- Directions for the design of future methods, platforms, and tools to support user-driven algorithm auditing, informed by our process model

2 RELATED WORK

2.1 Auditing for Harmful Algorithmic Behaviors

A growing body of work in human-computer interaction, machine learning, artificial intelligence, and other related communities has proposed tools, processes, and frameworks to audit algorithmic systems for biased, discriminatory, or otherwise harmful behaviors (e.g.,[22, 53, 62, 66]). Past research in this domain has uncovered harmful biases across a wide range of algorithmic systems like housing websites [9], hiring systems [21, 24], web search engines [59, 63], social media search [52], and e-commerce price steering [43] using different kinds of approaches led by experts and researchers. These approaches inspired a taxonomy to summarize the different algorithm auditing methods and research designs, including 1) code audits, 2) noninvasive user audits, 3) scraping audits, 4) sock puppet audits, and 5) crowdsourced/collaborative audits [66]. However, methods in this taxonomy often fail to surface serious issues that everyday users of algorithmic systems are quickly able to detect once a system is deployed in the wild [71].

Past literature has generally focused on expert-led algorithm auditing approaches that are performed outside the context of everyday use of an algorithmic system. However, this approach may fail when those conducting an audit lack the relevant cultural background and lived experience to recognize or know where to look for harmful behaviors [81]. In addition, certain algorithmic behaviors may only arise, or may only be recognized as harmful, when a system is used in the presence of particular real-world social or cultural dynamics that are challenging for external auditors to simulate or anticipate [46, 55, 69, 70]. Similarly, some harmful algorithmic behaviors may only emerge when a system is used in unanticipated ways or in unanticipated contexts [22, 35, 70]. As discussed next, recent research is beginning to explore ways to harness the power of situated users to overcome limitations of expert-led auditing approaches [71].

2.2 User-Driven Algorithmic Auditing

Recent years have seen many real-world cases in which *users* uncover and raise awareness about harmful algorithmic behaviors that they encounter through their day-to-day interactions with algorithmic systems, both individually and collectively [31, 44, 71].

In a recent, highly publicized example, Twitter users noticed problematic behavior from Twitter's image cropping algorithm and suspected that the algorithm was prioritizing white faces over Black ones (e.g., [14]). Twitter users came together organically to investigate. They tested Twitter's cropping of various images to assess whether the cropping algorithm exhibited a systematic, harmful bias against Black faces. They built upon each other's findings in online discussions to present evidence or counter-evidence for patterns discovered by other users or present new evidence and hypotheses of their own. Similar cases of "everyday algorithm auditing" have been observed around a range of algorithmic systems, including image search engines, online rating/review systems, machine translation, personalized advertising, and image captioning [71].

Researchers have begun to explore the design of systems that can harness the power of everyday users and crowds to surface harmful algorithmic behaviors that formal auditing approaches might otherwise miss [71]. A related line of work has developed prototype interfaces and crowdsourcing pipelines intended to support people in actively searching for algorithmic errors, biases, or harmful behaviors. For example, ImageNet Roulette was a simple online interface developed by artists and researchers to support users in exploring and interrogating the input/output space of an image captioning model trained on the ImageNet dataset. This project provoked discussions on social media, as users shared findings and hypotheses, and sometimes built upon each other's observations [19, 23]. In another project, Ochigame and Ye developed a web-based tool called Search Atlas, which allows users to explore and easily compare Google search results as if they were located in different countries [60]. Another body of work has explored the design of crowd pipelines, interactive visualizations, and interfaces to support crowdworkers in searching for and making sense of algorithmic errors (e.g., [10, 18, 73]). For example, Attenberg and Ipeirotis developed a game-like task called "Beat the Machine," which asks crowdworkers to find cases where an automatic predictive model is incorrect. While these efforts have demonstrated promising results, many open questions remain regarding how everyday users go about uncovering harmful algorithmic behavior, let alone how they are often able to be so effective at doing so.

2.3 Information Search & Sensemaking

In this paper we build upon existing work in the realm of information search and sensemaking to understand how people go about surfacing harmful algorithmic behaviors both individually and collectively. Pirolli and Card [61] proposed a general model of information search and sensemaking, based on cognitive task analyses conducted with data analysts. Their model posited four major phases of the sensemaking process, which may proceed through iterative cycles rather than in a linear sequence. These phases included information gathering, the representation of information in ways that aid analysis, the development of insights through manipulation of this representation, and the creation of some knowledge product or direct action based on these insights. Pirolli and Card's model organizes these phases into two major loops: the foraging loop and the sensemaking loop. At a high-level, these loops map onto the phenomena we investigate in this paper: how users search for and make sense of harmful algorithmic behaviors.

In the sensemaking literature, both search and sensemaking are understood to be *influenced by* the prior knowledge and beliefs of the sensemaker(s), as well as to *influence* their knowledge and beliefs [16, 47, 51, 65]. As such, we might expect that the strategies people use to search for and make sense of harmful algorithmic behaviors will be shaped, in part, by folk theories they hold about the workings of a given algorithmic system [26, 28] and by their existing perceptions and beliefs about algorithmic bias and unfairness in a given context [12, 15, 24, 38, 75, 80] — both of which are active areas of research in their own rights. However, it remains unknown how exactly such prior knowledge and beliefs inform people's search and sensemaking strategies in the context of everyday algorithm auditing, and it is unclear what other individual or group-level factors might influence these search and sensemaking processes. The current research aims to build on this prior research as a foundation to explore these questions, with the ultimate goal of informing novel design interventions to support more effective everyday algorithm audits in practice.

3 METHODS

In this study, we sought to investigate how people uncover harmful behaviors in everyday algorithmic systems: How do people find, make sense of, and evaluate potentially harmful algorithmic behavior? To explore this question, we conducted a three-phase study with 23 participants. We first conducted think-aloud interviews in which participants were tasked with actively searching for potentially harmful algorithmic behaviors online and making sense of these behaviors. These think-aloud interviews provided an opportunity to closely observe participants' thought processes and ask clarifying questions in a relatively controlled setting. Second, to complement these live observations, participants then took part in a 14-day diary study in which they documented harmful algorithmic behaviors that they found through active searching or during the course of their day-to-day interactions with algorithmic systems. The diary study enabled observation of participants' search and sensemaking strategies in a more naturalistic setting and over a longer period of time, allowing for the possibility of chance encounters with harmful algorithmic behaviors. Third, we then invited participants who completed the think-aloud interview and diary study to participate in group workshop sessions. In these sessions, participants were tasked with working together to evaluate and discuss some of the cases that were uncovered during the preceding diary study. We describe these phases in more detail in the following sections.

3.1 Recruitment

We recruited participants via social media (i.e., Twitter, Facebook groups, and Nextdoor), Craigslist Jobs, and mailing lists for local community organizations. Participant diversity was particularly important in this study as previous work suggests that lack of diversity in auditing teams may result in blindspots in finding harmful algorithmic behaviors [46, 71, 81]. Therefore, the recruitment blurb directed those interested to a demographic information questionnaire to provide characteristics including gender, race/ethnicity, sexual orientation, religion, political views, employment status, income, locations lived, languages spoken, level of education, tech savviness, and familiarity with algorithms. We pulled these characteristics from types of algorithmic bias in previous literature and framed relevant questions by drawing from a combination of other HCI research [68], the US Census survey [7], and a conference panel including Pew Research Center advice on categorizing religious groups [5]. Sixty-five people completed this screening form, from which we invited people spanning a diverse range of identity characteristics and backgrounds to participate. We first focused on inviting people from marginalized identity groups, as they tend to be underrepresented in research, then invited people with characteristics we lacked in our already signed up group. For example, after the first set of invites, we noticed a gender skew toward women in our signups, so we began to focus on recruiting men and non-binary people, inviting any who completed the questionnaire.

3.2 Participants

The screening process yielded a group of 23 participants (61% women, 30% men, 9% nonbinary) aged 35.2 on average with an age range from 20 to 64. Thirty-five percent of the participants were white, 30% Asian, 26% Black, 9% Hispanic or Latino, 9% American Indian or Alaska Native, and 4% Middle Eastern. (Participants could select multiple races/ethnicities.) See the appendix for more participant demographic details. Despite research recruitment commonly lacking enough people from marginalized identity groups, we noticed that we had somewhat less difficulty recruiting participants from marginalized groups (e.g. Black, LGBTQ) than expected. Additionally, no one who expressed interest in participating self-identified as "very conservative" politically, though that was a presented option, and we recruited half as many men as women participants despite inviting every man who filled out the demographic survey to participate. We speculate that these trends might arise from the topic's relevance to marginalized groups (and, conversely, diminished personal relevance to those who are not marginalized), mirroring recruitment patterns observed in prior studies related to algorithmic fairness and bias (e.g., [46, 55]).

The think-aloud interviews and diary study took place from November to December 2020, and the workshops occurred in February 2021. Participants were compensated with $20 for the think-aloud interview. In the diary study, participants received $5 per submission (up to a maximum of $30). Finally, participants received an additional $25 for participation in the workshop. We refer to the participants as P1–P23 throughout this paper. Unless mentioned otherwise, quotes are taken from think-aloud interviews.

3.3 Bias & Harm

As a study about bias and harm, our definitions and participants' understandings of these concepts are important in interpreting our results. In this study, we first asked participants about their understandings of these terms. This exercise revealed that participants did indeed tend to view bias and harm as distinct, but related concepts. Participants were then told that the current study was focused on *"possible cases of harmful bias and discrimination in algorithmic systems"*, which in turn was defined broadly in terms of instances where an algorithmic system *"behaves in ways that might cause harm for certain individuals, groups of people, or society."* We purposefully used this expansive definition to ensure that participants had enough of an understanding to complete the study tasks, while at the same time trying to avoid pushing participants toward overly restricted definitions. In the interest of investigating participants' perceptions, we wanted to avoid prescribing precisely what it means to be more or less harmful or biased.

3.4 Phase 1 – Think-Aloud Interview: Individual bias search & sensemaking

We conducted 23 remote think-aloud interview sessions, each approximately one-hour long, over Zoom. We chose Google Images as the domain for the think-aloud portions of our interview, given both that it was likely to be familiar to participants and because many historical cases of users finding harmful algorithmic behaviors have taken place in the context of image search [1, 4, 6, 39, 49]. For example, in a widely publicized case, users discovered that Google Images search results for the term "unprofessional hairstyles" overwhelmingly consisted of photos of Black women, compared with search results for the term "professional hairstyles" [2, 3].

The interview consisted of two steps. First, we asked participants to search for specific images. These image search tasks were chosen based on previous work that showed or research team discussions that suggested the image search results might be discriminatory or negatively impact some users. This step aimed to understand users' perceptions and evaluations of algorithmic behaviors anticipated to be biased in a harmful way. In the second step, participants started looking for new cases of algorithmic bias by exploring, generating, and trying out new ideas for search queries. This step aimed to aid our understanding of how users search for and make sense of potentially biased algorithmic behavior. Below we describe these steps in more detail.

3.4.1 Evaluating Existing Biases. At the beginning of the interview, we briefly described harmful bias and discrimination in algorithmic systems, then asked participants whether they had heard of this and, if so, what they thought it meant as well as any thoughts and reactions about times they might have experienced it in their own lives. We then asked participants to search for concepts that were known to be associated with societal biases across gender, race, sexual orientation, body image, religion, and culture, as based on prior work and research team investigations. We intentionally selected a diverse sample of tasks to try to avoid biasing participants toward certain kinds of biases and harms. This list of tasks included asking participants how they might find an image of a) a librarian, as the results showed mostly white women, b) a thug, as the results showed mostly Black men, c) one person nagging another, as the results showed mostly heterosexual couples where a woman nagged a man, d) a beautiful woman, as the results showed mostly thin, white women, and e) a wedding, as the results showed mostly Western culture weddings based in Christianity. We acknowledge that our perceptions of bias within these tasks may not represent the only biases present in them (indeed, participants identified others we missed); we used these categorizations merely as a method to select cases to use in the tasks. The cases appeared in the form of targeted think-aloud tasks that asked participants to search for images (using their choice of keywords), then interpret and discuss the search results.

Our motivation for starting with these targeted tasks was twofold. First, we wanted to see whether and how participants detected biases in the context of search queries that are known to have problematic results. Second, we used these initial, targeted tasks to get participants used to finding and making sense of harmful biases in the context of Google Images, in preparation for the less scaffolded, more open-ended tasks to come. After a participant completed a given targeted search task and interpreted the search results on their own, we probed 1) whether participants thought the search results were biased or discriminatory in harmful ways that might negatively impact people, 2) why (and how) or why not, and 3) what they would do if they had the option to change/adjust the search results.

3.4.2 Looking for New Biases. In the next portion of the interview, we wanted to learn how people go about actively seeking out new cases of harmful algorithmic behavior and what factors influence where and how they look (e.g., what search queries they decide to try, in the context of image search). Thus, participants were instructed to perform the more open-ended task of finding new examples of image search queries that might yield problematic image search results. As participants explored, generating and trying out new ideas for search queries, they were asked to think aloud to provide insight into their thought processes. Interviewers asked follow-up questions in order to better understand how participants generated ideas for search queries to try, why they searched for the specific terms they did, and how they interpreted the image search results they saw in response. At the end, we asked participants what they might do if they encountered something online like what they did in the session's tasks, what they thought should happen when they do an image search online, and for general session thoughts.

3.5 Phase 2 – Diary Study: Finding harmful algorithmic behavior in everyday life

We next invited interview participants to take part in a 14-day-long diary study in which participants were asked to look for potential cases of harmful algorithmic bias or discrimination online. The goal was to better understand how people encounter and make sense of harmful algorithmic behaviors on their own, either through their everyday interactions with algorithmic systems or by actively searching for such behaviors. Participants were asked to record observations of potentially harmful algorithmic behavior via a screenshot; a brief textual description; and an explanation of why they thought it might be harmful or problematic, whom it might harm, and how they found this case. 22 interview participants submitted reports, submitting just over 7 reports on average with the 4 as the lowest number of reports submitted by an individual and 15 as the highest. We encouraged participants to submit cases of harmful algorithmic behavior they weren't entirely certain about, giving them the freedom to explore their thoughts and suspicions in different situations without needing validation.

3.5.1 Interview & Diary Study Data Analysis. We conducted a thematic analysis approach of the data from both the think-aloud interviews and diary studies. The two interviewers (the first and second authors of this paper) performed an initial round of thematic coding by coding portions of the interviews and diary studies separately using NVivo. The goal was to capture a diverse and open range of qualitative observations, including reflections participants shared about how they conceptualize and understand bias and unfairness, the strategies they used to come up with search queries to find harmful algorithmic behaviors, and their perspectives on ways to remediate harmful algorithmic behaviors that they identified. Next, the research team collectively engaged in an iterative process of discussing and clustering codes into higher level themes. Below, we describe these themes.

3.6 Phase 3 – Workshops: Collective evaluation & validation of reported biases

Whether they arise incidentally as in the Twitter photo cropping case discussed earlier [44], or are initiated intentionally as in the case of ImageNet Roulette [19, 23], user-driven audits typically involve users coming together to evaluate and validate harmful algorithmic behaviors or potential biases that they or other users have observed [71]. Therefore, after investigating individual users' strategies to uncover harmful algorithmic behavior in the think-aloud interviews and diary studies, we engaged participants in collective follow-up activities. We conducted a set of four 60–90-minute workshops with 3–5 participants each, with the dual goals of a) understanding how people work together to evaluate user reports of potentially harmful algorithmic behaviors and b) beginning to understand how we might design to facilitate such collective activities.

In these workshops, participants were presented with de-identified reports of potentially harmful algorithmic behavior, generated and submitted by individual participants during the diary study. We curated this set of individually reported biases, drawing from the full set of diary study responses, in order to span a range of "types" of reported biases and harms. Specifically, we conducted a bottom-up tagging of cases submitted in the individual reports, with tags including gender, sexual orientation, race/ethnicity, socioeconomic/professional, age, body depiction, religion, and political focus in the reported issues. Given that our goal was to understand how users work together to evaluate potentially harmful algorithmic behaviors, we selected all submissions in which 1) the submission describes the issue as having to do with algorithms and 2) the submission describes the issue as harmful. This produced a set of 77 user reports. From this set, we randomly selected one submission with each tag (as submissions could have more than one tag, some overlap is present in this set). We also included a few user reports that did not fit the criteria above, motivated by knowing that such noisy reports do appear to arise in practice and wanting to see whether and how workshop participants might differentiate between these and reports of genuine algorithmic behaviors during their discussions. To get these, we randomly selected 3 submissions that did not follow one or both of the conditions above (i.e., one was not algorithmic, one was not harmful, and one was neither). Our entire selection process yielded 11 individual reports for use in the first workshop. After the first workshop, we chose a smaller set of 6 reports because reviewing this set of 11 reports took a substantial amount of time from participants in the first workshop.

Workshop Cases	Descriptions
Disability on Google Images	The participant searched "person with disability" on Google Images and the results were primarily about "visible disability" such as the use of canes and wheelchairs.
Pierced earrings on Google Images	The participant searched "pretty ear piercings" on Google Images and the results were all images of white women.
Game developers on YouTube	The participant searched "game design for beginners" on YouTube and the recommendation algorithms kept recommending related videos that target/represent men game developers.
Engineers on Google Images	The participant searched "engineers" on Google Images and the search results were predominantly about "civil engineers."

Table 1: Detailed descriptions of the four primary cases (user reports from our diary study) selected by workshop participants.

Since reports could have multiple tags, we removed reports such that we had one instance of each tag in this set of 6 reports.

We asked workshop participants to evaluate these individual user reports and work together to create a collective report that provides insights for someone who might be able to act on them. Each workshop proceeded in two phases. The first phase involved selecting a case to collectively write a report about. To do this, participants individually examined the user reports and advocated for what they thought was the highest priority to create a report for (e.g., in two workshops we asked participants to choose two of the submitted cases to focus on in the collective report and explain their rationales). Based on this, one user report was selected (e.g., based on which was chosen most by participants) as the focus for the next phase of the workshop. Table 1 shows the cases selected in this phase and discussed and written about in the next phase of the workshops.

In the second phase, participants worked together to make sense of the selected user report, asking and answering questions collectively about the reported issue. We provided a report template, designed to structure and prompt workshop participants' discussions. We generated the first, bare-bones version of this template based on our knowledge of the minimal information that might be important for decision-makers receiving such a report to know about an issue when attempting to address it. This template included prompts to get participants thinking about the desired report recipient, what happened, evidence available and evidence still needed, and actions moving forward. We iteratively honed the template over the course of the workshops based on our observations of how participants interpreted (or misinterpreted) each prompt. To help inform our broader project — toward designing ways to facilitate more effective everyday algorithm audits — we also iterated on other aspects of the workshop implementation across the four workshop sessions. For example, after noticing that participants took on some roles naturally and desired others that did not come as naturally, we began assigning roles in the second phase of the workshop. However, in this paper, we focus primarily on common themes that we observed across all four workshops.

3.6.1 Workshop Data Analysis. To analyze data from these workshops, our research team conducted a series of four Interpretation Sessions. In these, the research team first went over one workshop's data at a time to mark important observations as a group. Then, the research team iterated upon these observations through group discussions to cluster them into higher-level categories. Through our interpretation sessions, high-level themes emerged that were grouped with themes from the interviews and diary studies. We describe these themes below.

4 RESULTS

Our study results highlight the significant power that users have to surface potential biases and harms in the algorithmic systems they interact with. In the following sections, we explore this as well as a process model developed based on our analysis of the actions our study participants took and their explanations of theses actions.

4.1 Users' Power in Detecting Anticipated & Potential Biases & Harms

Participants showed great ability to detect and reason about potentially harmful algorithmic bias in both specific tasks where the research team anticipated bias might be found and in less prescriptive searches where participants directed their own investigations. In their evaluations, participants identified a wide variety of biases. Types of algorithmic biases named by participants included age bias, appearance bias, disability bias, gender bias, job type bias, location bias, cultural bias, political bias, racial bias, skin tone bias, sexual orientation bias, religious bias, and socioeconomic bias.

In terms of the ways that algorithmic behaviors could harm, participants described representational harm (i.e., lack of representation, overrepresentation, misrepresentation), the generation of negative emotions (e.g., offense, anger, annoyance, upsetedness), the perpetuation of incorrect ideas (e.g., misleading stereotypes), and sometimes no harm. Participants considered harm in a variety of ways: they mentioned specific communities or groups of people who would be negatively affected, such as marginalized identity groups and populations vulnerable to being misled (e.g., children and older adults were frequently mentioned); they connected to the potential for concrete physical or allocative repercussions; and they frequently pointed out that harmful algorithmic behaviors impact everyone, not just the most directly connected people, through the perpetuation of harmful ideas that damage society at large.

4.1.1 Existing Biases. In both the think-aloud interviews and the diary study, participants detected and made sense of a wide range of potentially harmful algorithmic behaviors. Participants were typically able to identify the kinds of biases that the research team had anticipated, and went further, pointing to additional potential biases in image search results. For example, while we gave the "wedding"

How this case was found (e.g., Where did you see this? What were you doing? How did you find/notice this? What is the context?):

“I used Google search for images of romantic couples. There images that came up were diverse, they appeared to represent couples from different cultures.”

Why this case might be harmful/problematic & whom it might harm:

“The only image for a biracial couple was a black and white hand holding no actual person, no images of middle aged or elderly couple, no same sex couple, except for a Hallmark movie preview of an upcoming movie featuring 2 males that would be highlighted in the movie. Middle aged, elderly couples of all races and cultures, same sex couples of all races and cultures could be harmed.”

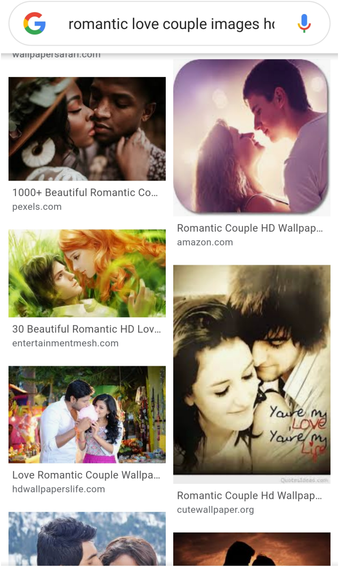

romantic love couple images h
wallpapersafari.com
1000+ Beautiful Romantic Co...
pexels.com
Romantic Couple HD Wallpap...
amazon.com
30 Beautiful Romantic HD Lov...
entertainmentmesh.com
Love Romantic Couple Wallpa...
hdwallpaperslife.com
You're my LOVE You're my Life
QuotesIdeas.com
Romantic Couple Hd Wallpap...
cutewallpaper.org

Figure 1: One example case submitted by our diary study participant P3 (“romantic couples”), including a screenshot of the potential bias, how this case was found, why this case might be harmful and whom it might harm.

search task due to its lack of representation of homosexual couples, many participants went beyond this to bring up other types of bias such as lack of representation of disabled people.

4.1.2 Potential Biases. In the next part of the interview, participants looked for new cases of bias and identified cases that were similar to the examples presented in the first part of the interview as well as entirely new observations of bias. For example, participants built on the “thug” search task to do similar searches such as “criminal” and “police” but also came up with searches like “Chinese food” that were unrelated to prior interview tasks.

In the diary study, participants submitted a total of 160 reports. 74 of these reports reflected observations that participants had made in the course of their everyday use of algorithmic systems, whereas the remaining 86 reports reflected discoveries participants made through active searching. The submissions described potential harm or bias spanning areas including gender, sexual orientation, race/ethnicity, socioeconomic class, age, body type, religion, politics, and combinations of these. 84 submissions included evidence or suggestion that the behaviors observed had to do with algorithms, and 126 included written evidence that the submitter thought the behavior was harmful. We will discuss many of the submitted cases in detail in the rest of the paper; see Figure 1 for an example.

4.2 Developing a Process Model of Bias Search & Sensemaking

Our analysis of data yielded a three-stage process model capturing the process by which people find and make sense of biased and harmful algorithmic behaviors as well as significant influences on different stages of this process. Figure 2 illustrates this model. In our open coding, three connected high-level stages of people’s bias search and sensemaking process emerged: search inspiration, sensemaking, and remediation. While the process tends to occur in this order, it is not necessarily linear, and each stage can occur after any other stage. *Search inspiration* encompasses ways in which participants came up with ideas for where to look for potential biases in an algorithmic system. *Sensemaking* involves the ways that participants understood and evaluated algorithmic behaviors that might be harmfully biased. *Remediation* covers participant actions and desires to mitigate the ramifications of harmful algorithmic bias.

Connected to each part of this process, we uncovered common influences on the process across participants: knowledge and beliefs, and platform affordances. Participants used prior *knowledge and beliefs* about biased and harmful algorithmic behavior to make sense of potentially harmful algorithmic behaviors they encountered, and they gained new knowledge and beliefs along the way. Participant capabilities were made possible and guided by *platform affordances* such as aggregate views in interface presentations and autocomplete suggestions by search engines. In the next sections, we first present a high-level description of the influences, then delve into reporting of different stages of the process.

4.3 Influences on Bias Search & Sensemaking

In this section, we describe major influences on the bias search and sensemaking process that we observed in the three phases of our study: 1) users’ knowledge and beliefs about bias and algorithmic systems and 2) the affordances for bias search and sensemaking that are present in a given algorithmic platform. In addition to influencing the bias search and sensemaking process, users’ knowledge and beliefs are themselves influenced by the process via the sensemaking stage. Below, we describe each of these influences and how they interact with users’ bias search and sensemaking process.

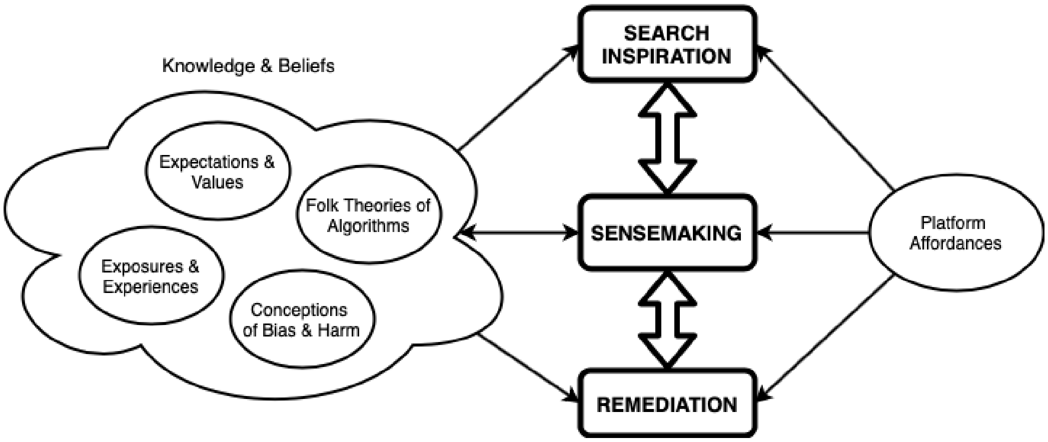


Figure 2: Process model depicting stages of and influences on users' bias search and sensemaking process. Bold boxes and text represent the three major stages of the process; other bubbles and text represent influences on and impacts of this process. Larger arrows represent movement between stages of the process, while the smaller arrows represent interactions between the process stages and influences on the process (i.e., platform affordances, user knowledge and beliefs).

4.3.1 *Knowledge & Beliefs.*

Almost all participants (n=22) revealed that they held pre-existing knowledge and beliefs about biased and harmful algorithmic behavior that influenced the search and sensemaking process. In Figure 2, a bidirectional arrow is shown between sensemaking and users' knowledge and beliefs, reflecting that these knowledge and beliefs both influence and are modified via their sensemaking of specific algorithmic behaviors. Participants referenced specific kinds of relevant knowledge and beliefs: a) conception of bias and harm, b) exposure and experience, c) expectations and values, and d) folk theories of algorithms, each of which we describe below.

Conceptions of Bias & Harm. The search and sensemaking process was heavily shaped by participants' conceptions of what constitutes "bias" as well as "harm". Participants defined algorithmic bias based on the intentions behind the design, the potential for imposition, and the priorities of algorithmic systems. Highlighting design intention, one participant defined bias as when algorithms *"don't actually do what they're intended to do"* (P7). Alternatively, focusing on the potential for imposition, another said, *"Bias means you are trying to influence the choice of the individual and try to impose your opinion"* (P18). Another participant defined bias by focusing on algorithmic prioritization, describing algorithmic bias as *"a preference towards certain individuals or certain languages or certain types of information, whether it's images or text, based on what is coded into the algorithm itself"* (P16). And some focused on the source of the bias, saying algorithmic bias arose via *"reflecting some human bias [...] injected through the data set or in the machine learning pipeline"* (P14) or via *"things that are calculated in the algorithm [...] through robotic systems that emulate humans"* (P22).

Participants' conceptions of harm also heavily shaped the search and sensemaking process. P3 mentioned a distinction between bias and harm, saying they *"see bias and negative impact and can't see harm."* In this vein, participants occasionally brought up cases that were algorithmic mistakes but not necessarily biased or harmful. For instance, although P14 would expect to see images of humans in the results of a search for "machine learning developer" (P14), the results mostly lacked visible faces and displayed abstract images and charts.

Some participants understood algorithmic harm based on the ramifications of algorithmic behaviors. A few participants saw harm as a primarily concrete and tangible effect. For example, participants described a situation where someone *"may not want to hire them"* (P16), which harms because *"it's a financial hurt"* (P10) and *"prevented someone from [...] employment, not being chosen, because of the way he dresses, the way he looks"* (P3). Others described how lack of representation *"further promotes the idea that you don't belong and perpetuates the sense of alienation"* (P2), how overrepresentation *"promotes a fairly narrow view"* (P12), and how misrepresentation *"reinforces a stereotype"* (P19). One participant teased out how harm might occur in some ways but not others: *"It's not going to cause any physical harm, but it could cause mental or emotional harm"* (P6). Another described a spectrum where harm is compared by severity: *"It doesn't really negatively impact someone to the extent that it's life threatening [...] harmful to a certain extent, but I wouldn't consider that superbly harmful"* (P21).

Other participants understood algorithmic harm based on the people affected by algorithmic behaviors. Some focused on those directly impacted. For instance, P1 described an image search for "thug" that showed predominantly Black men: *"It damages all of the Black community because if you're damaging Black men, then you're hurting Black families."* Others focused on groups of people less directly connected to the algorithmic outputs: P12 mentioned broadly that *"kids are especially open to being influenced from stuff like this."* Many described algorithmic harm as being *"problematic for everyone"* (P19) because anyone's *"beliefs and views would and could be shaped by this"* (P8).

Exposures & Experiences. Participants' prior exposures to and experiences of bias had large influence on how they went about searching for and making sense of biased or harmful algorithmic behaviors. In all the phases of the study, most participants (n=17) referred to a) their personal background and identity, b) their interactions with others, and c) external media sources when elaborating on the ways they had experienced and been exposed to biased and harmful behaviors. Below we describe these and their roles in the process of bias search and sensemaking.

Several participants (n=11) referenced their identity and the related experiences they had as members of particular demographic groups. For example, in the sensemaking stage, participants drew upon their individual backgrounds to consider how image search might manifest harmful and biased behaviors. While looking at the image results for "group of people", P1 said, *"From being a plus-size person myself, I think about body positivity and body neutrality a lot."* This helped P1 decide that lack of body size representation in the results had harmful repercussions. Participants also referenced their personal backgrounds and identities while in the search inspiration and remediation stages. For example, when searching "weddings", P12 said, *"I'm bi and [...] would love to see more couples of women represented."* In this way P12 connected their identity to imagined ways to remedy problems showing in the search results. Participants also brought up their relationships and interactions with others in the think-aloud interviews and diary studies (n=11). For example, P1 reasoned about potential harm in search results after being tasked with finding an image of a thug in the interview: *"My partner's Black, and I just hate this whole stereotype that the face of crime is a young Black man.* Similarly, in the workshops, participants often drew upon prior exposure and experiences to relate to the user reports presented. For example, when discussing how the presentation of disability on search engines might generate harmful results in Workshop 3, P18 quickly contextualized the user-reported issue within a specific usage scenario that they had personally experienced (the diagnosis of an autistic child in their family): *"We needed to search on Google or other websites, depending upon the symptoms. And we are not getting the correct reflection from the search engine so ultimately if someone is searching... it may be misleading"*.

News, information about current and past events, and other mass media informed people's exposures and experiences of bias as well (n=12). Beyond movies and television, some reported in the diary study that they *"decided to read current news"* (P4) in order to find potential biases and drew in the interview from what is printed *"in fiction [... or] in a newspaper or in a magazine"* (P14).

Expectations & Values. Most participants (n=20) anticipated how an algorithm would behave in a particular situation, forming expectations about algorithmic behavior. In addition to predictions of how an algorithm will behave, participants also expressed normative expectations, or thoughts about how an algorithm *should* behave, which they linked to their general values. Both what participants expected an algorithm *would* do and what participants thought an algorithm *should* do in a given situation were important parts of the process, as participants evaluated ways that an algorithm's observed behaviors diverged or aligned with those expectations. Participants began searching with expectations, then compared search results with their expectations to refine their conceptions of bias and harm (n=17). Sometimes participants expected and found bias (n=5). For instance, P12 conducted an image search for a "thug" and found overwhelmingly pictures of Black men: *"I've heard that the word 'thug' is generally racialized in addition to being kind of associated with delinquent or criminal activity, and so even just the history of that term kind of says– these results don't surprise me a ton."* On the other hand, sometimes participants expected bias but were surprised by search results (n=5). In the interview, P13 searched for "cooks" then "cooking" and said, *"I was expecting to see a bias towards females at home at least, which is obviously incorrect [...] yeah no, I think it's fine. It's not fitting any stereotypes."*

Folk Theories of Algorithms. Most participants (n=18) mentioned folk theories about why algorithms behave in the ways that they do, which informed their search and sensemaking process. For example, one participant thought that *"Google has no control over what is seen"* (P10), placing responsibility to algorithmic behaviors elsewhere [30]. Another participant considered certain images in a set of image search results: *"Who is searching them to the point that they're so high up that that's what people have to see immediately?"* (P16). In this way the participant put forth a belief that other users' behaviors are impacting the ordering of what shows up while also expressing dissatisfaction with the result [72]. Participants displayed a wide range of algorithmic folk theories that interacted with their expectations [27], values, and conceptions of bias [48] to influence how they conducted their searches [28], how they made sense of search results [29], and what ideas for remediation they generated [27, 76]. This corroborates previous work that users develop folk theories to understand and guide their behaviors within algorithmic systems [25, 26, 28][48].

4.3.2 Platform Affordances.

All three stages of participants' bias search and sensemaking process — search inspiration, sensemaking, and remediation — appeared to be influenced by platform affordances. For example, the Google Images search interface presents an aggregate view of multiple image results in a grid as its default view, which makes distributional trends (and deviations from these trends) easier to recognize: when searching for weddings in a specific task, P1 only mentioned that *"these all seem to be very Western-looking weddings"* after seeing *"one is actually a different culture."*. In other cases, Google Images' autocomplete suggestions inspired participants to search terms they might not have otherwise. One participant described a time when the suggestions caused them to be more precise in their search: *"As I was typing the word 'librarian', I could see a lot of the other choices that were popping up. And because I could see 'librarian memes', I didn't want memes, so that quickly caused me to modify my word"* (P19). Sometimes, Google's search query autocomplete suggestions inspired participants to propose related ideas to remedy issues they observed. For example, worried about being unknowingly influenced, P18 thought in *"that first initial searching, the user should be allowed to type completely through — no tooltip should be available."*

4.4 Stage 1: Search Inspiration

The first stage captures how people are inspired to search for potential biases and how people generate ideas for *where to look* for potential biases. This inspiration can occur incidentally, during a user's regular interactions with an algorithmic system, or it may occur intentionally, when a person is actively trying to come up with inputs to an algorithmic system (e.g., search queries) that surface biased or harmful behaviors. Two main themes related to search inspiration emerged through our analysis: noticing patterns (or lack thereof) and drawing upon prior exposures to societal biases.

4.4.1 Noticing Patterns (Or Lack Thereof). Sometimes participants noticed clear patterns or lack thereof in autocomplete suggestions, signalling the presence of potential biases and inspiring participants to look closer. For instance, P22 noticed when searching "nagging" that *"'wife' immediately comes up"* many times in the autocomplete suggestions. As another example, P23 noted during the diary study that Facebook stories have automatically suggested reactions saying "beautiful" that *"only popped up on stories where there was a woman or feminine figure featured or a pet."* Similarly, when participants noticed discrepancies between their expectations versus observed algorithmic patterns, they theorized about potential biases, and tried to look deeper. For example, P21 noticed and submitted a potential bias in the diary study *"because it was an advertisement for entertainment news and I do not follow any celebrities on my Twitter account"* and, similarly, P23 said, *"It struck me because it is so out of character for the folks I follow."*

4.4.2 Drawing Upon Prior Exposures To Societal Biases. Participants were often inspired to search for particular kinds of algorithmic biases based on their own prior experiences or exposures to societal biases. For example, when brainstorming during the open-ended interview task, P12 said, *"I study computer science, so I'm sure if you google computer scientists you would get a lot of men."* One participant explained that searching "image of a maid" was inspired by the racial bias depicted in the movie *The Help*: *"I thought I was gonna see some ladies from The Help [...] that was my first thought when I chose to search this particular image"* (P3). This participant drew from the movie's representation of Black women as maids in the movie. And P1 referred to a discussion with their partner about under-representation of women in the video game community as an inspiration: *"My partner really likes playing video games and we talked a lot about how the video game community skews male [...] so I'm going to try to look up gamer because that might pop up some things."*

4.5 Stage 2: Sensemaking

In the course of testing various inputs to algorithmic systems, participants inspected and made sense of algorithmic outputs to evaluate whether they might encode harmful biases. This sensemaking process generated new inspiration for searching, new beliefs about algorithmic systems, and new ideas for ways to remedy problematic algorithmic behaviors that participants observed. Below, we describe the related themes that emerged in the sensemaking stage.

4.5.1 "The Bias Is in the Eye of the Beholder": Interpreting algorithmic biases through the lens of their own biases. Participants made sense of potential algorithmic biases through the lens of their own stereotypes and prescriptive desires. The fact that *"everybody has their own definition about [...] bias"* (P4) led participants to subjective interpretations of whether what appears in search results is biased/harmful or not. In one case, many participants mentioned that the concept of "beauty" is subjective (n=6), so it is unclear how a search should represent that. For example, P1 said, *"I definitely do have more of that kind of Western society bias [...] I hadn't really thought about that kind of cultural aspect of it as much."* Similarly, when P10 was evaluating the search results of "beautiful woman" for potential biases, they did not consider the presentation of beauty via makeup a harmful bias because *"I feel like a woman should wear makeup. [...] I guess that's the way my mother feels, so it kind of drips down to me."*

Different conceptions of bias among participants sometimes prompted opposing views of whether an algorithmic output encoded harmful biases or not. For example, after searching "beautiful woman", P8 said, *"I was pretty sure [...] everybody would probably be represented here"* and concluded that *"everybody like I said is represented here as far as race goes."* However, many other participants commented on the lack of racial representation when they performed the same search task: *"They do need more diversity [...] I don't see anybody with different skin tones"* (P7).

4.5.2 The Interaction between Bias Evaluation & Perceptions of Reality. Participants compared algorithmic outputs with their perceptions of reality to evaluate what is and is not biased (n=13). For instance, P13 considered the results from a specific task to find an image of a wedding and said, *"As an Indian, I know that it's not accurate for India"*, comparing their experience of reality to help them examine potential cultural bias in the search result. Some participants believed that results that reflected reality were necessarily less biased: *"the demographics fit the population which is what I think it should come out to be when you do a search."* (P5). In practice, this led to the expectation that search results should reflect reality: For example, P5 thought a search for "professor" during the interview might not be biased despite showing little gender and racial diversity because *"maybe most professors are white males, in which case I think it would be fair that's what the search shows"*, or similarly, *"If all firefighters were white men, then that's I think what should come up"* (P5).

Some had more nuanced insights when comparing with reality, saying that, though a search might technically be accurate to real-world data, *"the data might not be reflecting the reality"* (P14). P14 provided an example: *"If you look at the surface there are indeed more African Americans being arrested or reporting violent [...] but then this is institutional bias, right?"* In this way the participant elaborated that data is not a neutral reflection of reality and can contain biases as well. Similarly, some thought the results were *"maybe not necessarily inaccurate towards how [a search term]'s used. It's just, the way it's used itself is kind of biased"* (P12). In line with findings from [46], a few participants noted that some biases might reflect reality in informative ways, and preferred to preserve these biases: *"It's reflecting human condition and human bias, which is actually good because if you make it extremely balanced, it's almost unreal"* (P14) and *"it would seem almost wrong if it was all reversed"* (P5). However, it has been pointed out that real-world search results can conflict with outcomes desirable to society [49].

While participants' perceptions of reality impacted how they interpreted and evaluated potential biases of algorithmic outputs, they were also shaped by algorithmic outputs. A few participants (n=7) who believed representation in search results should reflect reality concluded that the search results were likely accurate *regardless of any biases present.* In line with prior literature on perceptions of bias in image search (e.g., [49]), these participants then used search results as a basis for making inferences about reality. For instance, P10 said about a search, *"I see a lot of women, not a lot of men"* and concluded, *"I guess it's a predominantly women's field."*

4.5.3 Reflecting on Uncertainty in Bias Evaluation. While participants tried to interpret the algorithmic outputs in various cases, several (n=10) described difficulties in knowing whether biases were actually present. Repeatedly, participants mentioned that they *"really didn't have any solid evidence, it was just a feeling"* (P3). Participants pointed out that the information available to them was insufficient to prove some of what they noticed. For example, when looking at image search results for "beautiful woman", P19 said, *"I also don't see anyone who would jump out to me as a transgender woman, and so I do believe that this leaves them out. And I mean, sometimes it's hard to tell."* In this way, P19 detected a potential lack of representation while simultaneously acknowledging that the way someone presents does not necessarily have bearing on that person's gender.

Reflecting on their uncertainty, participants discussed the information they would need to have a more certain evaluation of an algorithmic outcome. For example, in the course of their collective investigations in the workshops, participants often reflected on missing contextual information needed to properly evaluate a given user report. In some cases, participants recognized that they lacked crucial information needed to determine whether a reported issue represented biased and/or harmful algorithmic behavior. For example, in Workshop 1, when discussing the pierced earrings case (refer to Table 1), P5 expressed uncertainty that the reported behavior actually resulted from algorithmic decision-making: *"I don't know [...] I was under the impression it was an ad from a company, you know it wasn't a Google search, so I just thought, maybe that company sells that specific style of earrings (for women) [...] men wouldn't wear them."*

Workshop participants sometimes asked specific relevant questions about information they lacked for a reported case. For example, in the Workshop 4 group report, when trying to assess whether a user report provided evidence for a possible gender bias in the YouTube recommendation algorithm, participants reflected on the missing information needed for an effective investigation: *"Does the user have personalized ads on? Their actual gender and age? Incognito mode or not? Prior search history do they often search for this kind of content?".*

4.5.4 Advocacy & Solidarity. As participants reasoned about algorithmic harms, they sometimes advocated for marginalized demographic groups. . For example, during Workshop 1, as participants discussed the limited diversity in search engines' representations of romantic couples, P10, middle-aged at 46, argued that elderly couples would be harmed: *"When my grandma said she had a boyfriend, [my little cousin] was like, 'Grandmas don't have boyfriends.' [...] It's just a stereotype and it's ingrained in people's memory."* :While advocating for marginalized groups, participants sometimes discussed the further consequences of such algorithmic biases: P16 in Workshop 2 said, *"By whitewashing these disabilities, whether they're visible or not, it also leads to some of the issues in the medical field where doctors don't trust women and also minorities and people of color when they talk about their pain threshold."*

Participants' advocacy of marginalized groups sometimes turned into collective solidarity during workshops when participants worked together. For example, we observed that by drawing upon prior experiences and invoking parallel cases, participants sometimes attempted to form a shared identity with others in their workshop groups. A particularly interesting example emerged during Workshop 1. Although the group consisted of participants with different self-reported demographics, including two women and one non-binary participant, participants seemed to form a shared femininity in their discussion of the pierced earrings case. For instance, in the Workshop 1 final report, the group elaborated on why this was harmful: *"Society dictates how we act as women and how we should dress and wear jewelry. It is kind of expected by others that we wear these things in public."*

4.5.5 Nuances in Meaning of the Selected Algorithmic Input. Participants also considered the various ways their inputs to algorithmic systems could be interpreted when considering potential bias (n=7). Some noticed linguistic connotations because search results depicted a singular version out of many possibilities: P2 image searched "elegant" in the diary study and said, *"It almost exclusively depicts very thin, white women in expensive evening wear [...] 'elegant' can just as easily apply to a math concept or sculpture."* On the other hand, some participants noticed linguistic elements because search results depicted more possibilities than what was imagined while searching: P11 image searched "model" in the diary study and said, *"interestingly, one of the results is a biomedical model of the skin [...] I didn't stop to think about other semantic interpretations of the word 'model'."* This could lead to participant analysis of word meanings, as participants considered how definitional and interpretational differences between terms might affect a search. For example, when image searching terms like "fashion", "cute dress", and "sexy" in the diary study, P17 found mainly images of white women and said, *"These words even though they are not gendered seem to have heavy gendered interpretation."* Participants wondered about how some words can have more neutral connotations, how other words *"are imbued with a harmful, [racist] connotation, and how this arises in images online"* (P11). Some mused about how communities reclaim certain words used negatively to describe them and what that means for image search representations. For example, P4 described how "thug" could be highly associated with Black men in search image results partially due to some Black men rappers choosing to self-describe as "thugs". P19 mentioned that "thug" could have been reclaimed in the same way as "queer", *"a word that has been reclaimed and chosen by the LGBTQ community".*

4.5.6 Sensemaking through Comparison. Many participants used comparison on the platform as a method for evaluating and making sense of search results (n=14). Sometimes this meant doing more searches as a sort of test. For example, P16 image searched "good neighborhood" and then, in order to make sense of the search results, searched "bad neighborhood", comparing the visual depictions of the two. Similarly, participants followed up image searches for "nagging" with searches for "men nagging other men", "nagging man", and "man nagging" in (unsuccessful) attempts to force depictions of men who are nagging into the search results (P10, P14).

Other times, this meant comparing the amount and type of representation. In Google Images, this could take place across multiple searches, as participants used the amount of representation in one search result to situate themselves in another, or in a single search, as participants compared resultant images to each other, something Google Images' search interface makes straightforward. Perplexingly, noticing one or two images of a type of person sometimes led participants to conclude that the ratio of representation seemed bad within the set of results and thus the search was biased, but other times led participants to conclude that various types of people were represented and thus the search was not biased. Past work has discussed the role of comparison in sensemaking and other cognitive processes [36].

4.5.7 Placing Responsibility for Algorithmic Outputs. Participants considered what controlled platforms and caused bias to appear. Some participants believed that Google intentionally produced the representation — or lack thereof — in image search results (n=4). When more diversity appeared in a search result, participants expressed that *"they're trying to achieve better representation"* (P12). Participants left "they" unspecified, a pronoun with no antecedent, some person or people who have the power to unleash their intent in Google Images search results. When less diversity appeared, participants similarly claimed that *"there's a pretty narrow definition being pushed here"* (P12).

Others theorized that the algorithm itself was neutral and that results were a reflection of the data (n=6). Explaining the lack of diversity in a search result, P10 said, *"They might not have the picture available"*, moving responsibility from the algorithm to the data. Others posited that Google was *"not hand picking this stuff"* (P12), blaming their own search behavior for their inability to find better representation in a search: *"It's probably because I'm not searching [...] this is just a brief search. If I took the time, I'd probably find some"* (P4). This suggested that search result bias is symptomatic of the searcher's bias in selecting certain search terms in which they could *"demonstrate something without thinking"* and inadvertently neglect *"other things that [they] could choose instead"* (P11). Others blamed results not on individual users but on users en masse, claiming that results were *"based on what people have been typing into the search bar"* (P21).

Many participants thought the algorithm targeted individual users (n=8). Participants described this as personalization that provides *"the response that's best suited to the question you're asking"* (P16). Some participants thought this happened via *"targeting advertising based on the metrics that I provide"* (P23), leading to a situation where, for example, *"they think I'm a 40-year-old man because of my athletic preferences or something because I clicked on some shoe thing"* (P11). Another theory suggested that the search results changed based on time relevance: *"I think it'd be interesting to see how the algorithm changes with current events because it's definitely something that would change on a day-to-day basis"* (P11). Many hypothesized that results were *"probably based on [their] location"* (P1) since *"Google definitely knows where I am"* (P13), which led participants to *"wonder how this changes when you're looking at [results] from different countries"* (P16).

We similarly observed in the workshops that, in the process of collectively investigating whether an individual user report represented a case of harmful algorithmic behavior, participants questioned how the behavior originated. However, they often had a hard time untangling different possible explanations for an observed behavior based on the information provided in the user report. For example, in Workshop 2, when discussing the disability case, participants started a discussion to figure out the sources of bias and eventually concluded that it could be a mixture of multiple factors: *"It could be algorithm or human... or sometimes it's both? Even if there was a developer, putting in the source code in the beginning, if input [is biased...] the computer will be running on that additional information"* (P16). Responding, P21 said: *"It's a couple of different factors. Like how much people are searching, that could be one of the factors"* (P21).

4.6 Stage 3: Remediation

The third stage describes the actions that people take to avoid or counter harmful biases. We observed that, while reflecting on potential changes that could be made at a platform or societal level to counter harmful biases, people often reinterpreted the algorithmic behaviors they had observed, questioned whether these truly represented biased algorithmic behaviors, and reflected on relevant information that they lacked, leading them back to sensemaking. Through sensemaking, some remediation actions also informed search inspiration. For example, in the context of image search, we observed that some people avoid searching for a particular term or alter their search terms in efforts to avoid seeing potentially harmful biased search results.

Participants valued remediation of potentially biased and harmful algorithmic behaviors (n=22). Through their actions and desires, they expressed ways to try to fix potential issues they encountered. When they were able, participants took action; other times, participants described methods for the platform or society more broadly to take action. Five types of remediation emerged — representation, resistance, awareness, pressure, and communication — and are described below.

4.6.1 Improving Representation in Content, Diversity, & Ordering. Improving representation included desires for increased diversity and for increased accuracy in representation (n=17) — that is, wanting *"to get accurate results and [wanting] to get diversified results and [not wanting] to use stereotypes or certain biases against certain types of people"* (P22) — goals that are sometimes, but not always, aligned. This tension highlights that bias removal is not always a reasonable option, as people have different values, expectations, and conceptions of bias. Beyond broadly thinking that results *"should represent a more diverse group"* (P17) or should *"be more statistically correct"* (P13), participants described concrete methods of curation by which they thought these representation aims could be achieved.

One described method involved bringing some of the images *"up to the top"* because *"that might be more balanced than what pops all the way to the top"* (P19). Reordering images to *"add in more [diversity] in the first two rows"* (P21) would ensure that representation appears *"in the beginning, where people most are paying attention"* instead of further down the page, *"where people aren't going to be scrolling"* (P9). A related method put forth that *"it'd be good to have* *one [image] after the other kind of like alternating with different faces, different ages, different ethnicities and races, just to make it clear that everyone's represented here"* (P16). To achieve this curation, some pointed to a platform-level need to *"be more careful about balanced data set"* (P14) and even consider *"[doing] it in a way where [platforms] have their own images"* (P10) instead of pulling from other sources for data. With or without curated results, P18 emphasized that users can always *"go more deep to the search or go to the next page"* to get more representation.

4.6.2 Individual User Resistance. Participants took a variety of actions to counteract bias themselves (n=12). Some created representation themselves out of inadequate results by *"finding images [...] that don't perpetuate those stereotypes quite as much"* (P12). When potentially biased results appeared after using a search term, others opted *"to not search for that [term] and try to search for something else"* (P7), to *"try different word searches in the search box"* (P4), and to *"be more specific in the way I search"* (P16). For example, after noticing a dearth of cultural representation in a "wedding" image search, P13 said, *"I'd type in 'Indian wedding' and [that] will definitely be what I would be expecting for sure."*

For some participants, being reactive was not enough: they sought preemptive remediation. When tasked with finding an image of a wedding, P11 searched "lesbian wedding", explaining, *"I'm in a lesbian partnership right now and wanting to get married and envisioning a wedding [...] and I'm so sick of [seeing] these straight weddings."* Similarly, when tasked with finding an image of a beautiful woman, P19 searched "women", explaining, *"I was concerned that if I typed in 'beautiful women' I might not like it. I might get a whole lot more white people, and I might get more swimsuit models and stuff like that."* In this way, participants showcased their awareness of biased norms, highlighting that at some level they knew they needed to alter their search terms to try to guard against this and get the desired results.

Other participants took even stronger measures. Some attempted to prevent their ability to see potentially harmful content by trying *"to block out the stuff that offended me"* (P10) and *"to avoid that thing, block that particular person or group or website"* (P18). And one participant said bias could make them*"reconsider the kind of mediums that I'm using, the kind of information I'm taking in"* (P16). Such attempts to resist algorithmic bias align with the concept of everyday algorithmic resistance [76] in which users try to exercise their agency over algorithmic structures. Past work also showed users resisting algorithms to define their own identities in relation to the algorithm [48].

4.6.3 Fostering Awareness of Bias. Participants sought to bring bias to the surface of people's minds through raising various kinds of educational awareness (n=16). Sometimes this awareness raising occurred at a societal level. For example, P14 said that a search result was *"so bad that it's actually good for society because it just reminds people how this kind of stuff has happened before [...] we should really dive into this phenomenon and look into the root of the discrimination and try to change it."* In this way, P14 suggested that demonstrations of bias could spur further anti-bias work.

Some thought that users should be *"trained how to use the technology [...] what are the do's and don'ts, what are the ethics"* (P18) and taught *"how to do more robust image searches and more robust searches"* (P19). Some particularly suggested public education. For example, in Workshop 2, when discussing how certain forms of disability (e.g., depression) are rendered invisible in image search results, participants advocated for public education around the nature and scope of disability as a path toward addressing this: *"They [the public] need to be educated about what a disability is"* (P19). Similarly, participants in Workshop 3 wrote in their group report that *"A lot of regulations around [accessibility]... people are just not aware of it. There are a lot of organizations to go for. It's well researched and documented but not getting enough attention from the population. [We need to] survey the public about their awareness."*

Additionally, participants thought that the tasks they were asked to do as part of the think-aloud interviews and diary study could be useful in educational contexts. P11 described being able to *"use [these] as teaching tools a lot. I teach a lot about harmful stereotypes [...] like we do semantic prototypes for one of our topics,"* and P12 thought that the tasks were *"super valuable in terms of portraying [algorithmic bias] pretty clearly"* and could be useful to *"teach about algorithmic bias in some kind of interactive thing."* Participants felt the tasks were educational for themselves as well: *"It was very important, revealing just the disparaging of minorities on a daily basis, in certain places"* (P8).

Some participants said that they would raise awareness on a more personal level. Telling others who think and feel similarly presented as a method to both raise awareness and share feelings. For example, P1 said, *"I feel like my friends and I will commiserate on it, since it's something that several of us experience."* Similarly, P22 would *"share with my friends to kind of laugh at why is this the first thing that shows up?"* Telling others who might not think and feel similarly also came up. One participant did this *"to see what their opinions are"* (P9). Another *"[raised] these points to my peers who may not be as aware to start the conversation"* (P16). And others described doing this to teach others who *"could probably benefit from hearing that discussion"* (P1). For example, P12 said, *"I've got a little sister. I would probably be like, 'Yo, these search results are kind of bad. Here's why.'"*

4.6.4 Exerting Pressure on Companies. During our workshops, participants expressed desires for companies to take responsibility and fix issues on their platforms. Collectively, they described how they might pressure companies to take action using the aforementioned awareness raising. For example, in Workshop 4, P13 wanted to give the report to *"a news agency, because I think if it went straight to the developers they might ignore it if it's something that makes them money."* In this way, raising awareness connects to influencing a company via publishing issues to a wide audience. As P9 also in Workshop 4 said, *"You want the news to get out to the whole public and reach a wide variety of viewers."* Similarly, P16 in Workshop 2 would *"recommend reaching out to civil society and maybe nonprofits that [...] would have a stake in this and sometimes they can put more pressure on larger companies like Google."* As a concrete way to exert pressure, P19 in Workshop 2 brought up money, saying they *"could also pick a couple advertisers who would not want to be associated with those biases and see if we could get their attention on it, [see] if we couldn't get Google to act."*

4.6.5 Communication between User & Platform. Participants thought platforms could communicate with and understand users in ways that better address harmful algorithmic behaviors (n=15). A frequently mentioned method to accomplish this involved changing the communicative burden [54] to rely less on user behaviors. Some participants wanted the platform to understand users via idealized personalization that could *"magically come up with the thing that I want in my head"* (P11) so that users *"get the response that's best suited to the question you're asking"* (P16). However, participants knew they could be *"so biased by everything that it pops up"* (P19); because of this, one remedy proposed that *"the search engine should allow the user to type the complete criteria or word or sentence [they] want to search"* (P18).

Others wanted the platform to provide additional information about what was happening. Some thought explanations of algorithmic behavior and how to interact with it could help: P21 thought *"a note [...] explaining why certain searches give you certain results could be useful,"* and P18 similarly said that platforms *"should provide a module or instructions, do's and don'ts on the top."* Though this would be done by platforms, it would presumably result in changed user behavior that leads to less biased results. Others wanted platforms to provide more context about some searches, such as by offering users *"all of the different meanings [of a term] up in the first screen"* (P19) or through *"a feature that allowed you to get a description of the term before you see images"* (P16). Another mentioned way to include context was through *"some kind of disclaimer saying, 'This word has a racialized history'"* (P12) in relevant places.

Participants also would appreciate increased ability to give feedback to platforms. Wanting to report a concern, P3 wondered, *"Who do I write to, to make this complaint?"* In that vein, P18 said, *"If [the] option is available, then I'll submit [issues] as spam or harmful post."* Others described a slightly different interaction style where *"you should be able to flag [issues] for someone to either remove or add context"* (P16) and *"mark them as inappropriate"* (P13). Participants emphasized the need for ease, as they are *"not going to go out of my way to look for people to tell"* (P8), and for action to be taken in response, as P11 described situations — *"when I tend to report harmful things on Facebook or Instagram [...] it still doesn't go anywhere"* — that led to frustration with ineffective reporting processes.

5 LIMITATIONS

It is important to highlight some specific limitations of our methods. In this study, we focus on image search as our problem domain for the interview study, as it provides a familiar, straightforward starting point for our participants to search for problematic machine behaviors. It offers an interface that supports easy comparison across different images, through which biases in result ranking and distribution can be more easily identified. However, biases are also present in other types of algorithmic systems and digital platforms, where such patterns may not be salient. In addition, this kind of comparison-based pattern detection is not always relevant to identifying harmful machine behavior, which can manifest as a single problematic instance (e.g., an image-caption pair) rather than a statistical pattern [40]. Also, when considering harmful bias, a large number of participants brought up biases that were caused by misrepresentation and stereotypes. This focus might have been influenced, at least in part, by the examples we chose in the specific interview tasks and the use of image search as our initial problem domain in this study.

6 DISCUSSION

Past literature has shown that everyday users can be powerful in detecting harmful and biased machine behaviors in their daily interactions with algorithmic systems [71]. However, little is known regarding how such power can be harnessed and guided. In this paper, we have begun to shed light on these questions by investigating the strategies everyday users use to surface potentially harmful algorithmic behavior, both individually and collectively. Below, we discuss design implications and future research directions toward supporting user-driven algorithm auditing.

6.1 Supporting User-Driven Algorithm Auditing

In this paper, we investigated how people surface harmful algorithmic behaviors with the aim of informing the design of new tools and methods to support user-driven algorithm auditing. We presented a process model capturing influences and stages of users' bias search and sensemaking process. Our investigation primarily focused on scenarios in which participants *actively* search for potential algorithmic biases; future investigations could explore how we might better support people in following up on incidental encounters with algorithmic biases. Here we discuss how our process model can inform the design of new interventions to facilitate future user-driven audits.

6.1.1 Utilizing Prior Experiences & Exposures in Uncovering Harm. The results of our studies suggest that people's exposures and experiences have a large influence on what types of bias they are able to identify. In addition, beyond their personal identities and backgrounds, participants drew upon second-hand knowledge of the experiences of those close to them, or knowledge from media that they had consumed. Assigning people to particular auditing roles based on their specific exposure and experience could boost future auditing efforts. For example, in order to surface potential racial biases in an algorithmic system, it may be most useful to engage users with exposure to and experience with related racial bias. Further, this indicates that having auditors with a diverse set of exposure and experience is crucial to comprehensively unearth harmful algorithmic behaviors.

Participants also showcased their own cognitive biases as they made sense of potential biases. Sometimes participants entered a term and said that they searched it because they expected a certain bias to appear. Then, despite having little to no evidence to support the expectation, they concluded that their expectation of bias was correct. It may be that people held preconceptions so strongly that they fell prey to confirmation bias. Future supports for user-driven algorithm auditing should anticipate such cognitive biases, and should be explicitly designed to mitigate their impacts.

6.1.2 Facilitating Bias Detection & Reporting via Platform Affordances. In our studies participants mentioned desires and actions to raise awareness of problematic algorithmic behaviors. Participants described telling other people that they know, sharing information with impacted groups, and getting the word out via mass media, and they also expressed that they would like to communicate issues with platforms and companies. We can leverage our knowledge that platform affordances influence the process by providing straightforward ways to report bias, either to the platform or beyond, through the design and mechanisms of the platform itself.

We also know that participants frequently used comparison as a method by which to notice, understand, and evaluate potential bias. For instance, Google Images' search interface allows users to easily view groups of *multiple* images in ranked search results, which is useful for examining patterns and finding biases that only emerge through comparison. Participants also used comparison to look for more evidence, testing for the presence of bias through levelled searches that they would then compare to each other. This proved helpful to participants' consideration of bias. Employing platform affordances to facilitate this comparison could lead to richer and more fruitful user-driven algorithm audits.

Another type of platform affordance comes in the form of quantification. Although we see potential for user-driven auditing in identifying qualitative harms, participants sometimes attended to quantitative information present in the platform in forming their judgments. As one example, participants often looked for proportions represented in image search results, which could be productive in surfacing certain issues where participants noticed a dearth of certain categories. But we also observed several cases where we believe it was distracting to look at proportions, such as when participants identified the presence of a single instance of a category as proof that there was enough representation. So quantitative reasoning for qualitative harms can be helpful or counterproductive, depending on the method.

6.1.3 Toward Building Algorithmic Counterpublics. As previous literature [71] suggests, when a group of users work together to collaboratively make sense of problematic algorithmic behaviors, they often form what Nancy Fraser termed as "counterpublics" [34] – where members of often marginalized social groups collectively participate in their own form of sensemaking, opinion formation, and consensus building. For example, via the platform's internal discussion forum, Yelp users were able to form algorithmic counterpublics against biased rating algorithms [33].

We observed similar counterpublics forming in our four workshops. Indeed, in the workshop sessions, we saw how participants from various backgrounds attempted to establish common understanding around the issues being reported via actions such as untangling the sources, reflecting on the context, and proposing possible remediation solutions. How can we better support such collective auditing behaviors to build more successful algorithmic counterpublics?

We observed that the lack of contextual information remains as one of the major challenges for our workshop participants, as they often had a hard time untangling different possible explanations for an observed behavior based on the information provided in the user report. One way designers might help auditors overcome this challenge is to allow them to ask follow-up questions of people who report their observations. For example, since none of these workshops had access to the demographic information of the user report submitters, participants were constantly questioning whether the problematic machine behavior could be a type of "demographic targeting". Future designers should consider strongly encouraging auditors to provide necessary contextual information (e.g., demographic information) to better facilitate such collective sensemaking. Another way for designers to support such types of collective sensemaking is to offer cues to remind the auditors to reflect on the missing information. For instance, in workshop four, participants did not necessarily reflect on missing information until filling out the final report, after they had already discussed a given case at length. In this case, some of the questions in our workshop's final report served as cues for our participants.

We also observed interesting dynamics of the demographic makeup of our workshop participants. For example, if in a workshop, all participants shared similar backgrounds (e.g., shared experience or similar technical expertise), they might be missing certain information in bias validation due to the constraints of team diversity. However, sometimes similar backgrounds and shared identities might also allow people to bring in new ideas without them seeming threatened. Future designers should consider how to balance such dynamics in supporting more effective team formation.

There is great opportunity in encouraging users to come together within their auditing activities. Though users can and do individually audit algorithms, oftentimes it is only together that can they conduct tests as different, authentic users to gather robust evidence. Only together can they discuss and provide general consensus or disagreement, or raise awareness at the scale required to put pressure on companies and hold them accountable.

6.2 Limits & Challenges of User-Driven Algorithm Auditing

Though user-driven algorithm auditing has many strengths, it comes with its own challenges and limitations. Everyday algorithm audits have had varying levels of success [71]. One relevant factor may be the difficulty in creating productive structure and organization in a user-driven algorithm audit. Establishing structure and organization is especially difficult amongst large groups of users, though collective action is exactly what many user-driven audits need to interrogate, hypothesize, test, and raise awareness effectively. Our process model offers insights for how users work, together and independently, that can be leveraged to carefully create and support this structure and organization at scale.

User-driven audits thus far have occurred in situations where users notice potential issues within algorithmic systems, whether by stumbling upon or by actively searching for problematic algorithmic behaviors. This limits the relevance of user-driven auditing to algorithms that are directly visible to users to interact with, neglecting many algorithmic systems that are offline and invisible to users. However, tools and platforms that facilitate user-driven auditing can be designed to provide and encourage interaction with algorithms that users do not normally interface with. A similar but distinct challenge involves situations in which those targeted and potentially harmed by algorithms are not the the direct users who interact with these systems day-to-day. For example, in many high-stakes algorithm-assisted decision-making contexts like child welfare decision-making and recidivism prediction, public sector workers are the direct users of such systems, but the citizens and usually marginalized people are the ones who are impacted by these algorithmic systems' outcomes. In such settings, it may be beneficial to bring in diverse stakeholders with relevant knowledge—especially those who are impacted by algorithmic outcomes.

Not all algorithmic behaviors are equally visible, and sometimes people are harmed by algorithmic behaviors without their knowledge. In this paper we have largely focused on harmful algorithmic behaviors that people have opportunities to encounter in their day-to-day lives and that (at least some) people can plausibly be expected to recognize as harmful. Not all algorithmic harms fit this description. In some cases, this may be because those most affected by a given algorithmic system are not the users who directly interact with the system, as discussed above. It may also be due to the opaque nature of some algorithmic systems. For example, gig workers often are unable to tell how the algorithm managing them works. But apps to compare wages have helped gig workers to collectively reveal inequities [57]. The creation of new platforms like these apps, with affordances that support collective investigation and sensemaking, can provide avenues for the recognition of harmful algorithmic behaviors that may otherwise remain invisible.

Determining what is noise and what is legitimate, relevant information is another challenge. Often the users are not technical experts and have no straightforward way beyond their own input/output based testing to determine the correctness of their hypotheses about how the algorithms work. Identifying signal versus noise is recognized as a challenging problem for user- and crowd-driven approaches in general. Some bug bounty programs use experts to ensure they have more signal than noise [37]. But weeding out noise requires care, as the process can involve a level of subjectivity. Apparent "noise" may actually be an important signal, indicating where the ground truth is contested (e.g., representing substantive disagreements across different social groups) [56, 67].

Of note, *supporting* user-driven auditing inevitably involves priming users' perceptions and auditing behaviors to some degree. Just as we primed participants with specific tasks and examples in this study, users on a platform would also necessarily have to be primed to some extent in order to support the process of user-driven auditing. For example, the onboarding or training provided to users who are new to a user-driven auditing platform may influence their subsequent bias search and sensemaking behaviors, no matter how diligent platform designers are in reducing such priming effects. Similarly, if a user-driven auditing platform recommends specific auditing tasks to users, there are risks that the selection and presentation of tasks will inadvertently serve to prioritize the interests of some social groups over others, or that the structure of a given task will facilitate identifying certain kinds of biases over others.

6.3 Algorithmic Harm Remediation: What's next?

A critical question that any user-driven auditing platform must be designed to address is this: in cases where users identify harmful algorithmic biases, what might it actually look like to "fix" the issue? Our participants showcased varied definitions of bias through their differing interpretations of algorithmic outputs. Some participants envisioned accuracy and matching with reality, while others envisioned diverse representation regardless of accuracy. For accuracy, the process of choosing ground truth is fraught, as many sources are littered with biases and other issues themselves. For diverse representation, what would ideal representation look like? Interestingly, participants found representation important in both positive (e.g. "beautiful woman") and negative (e.g. "nagging") contexts. This begs an ethical question about when the negative context is harming those who are represented and thus increasing representation would consequently harm more types of people: Is representation the answer? And if it is, who or what should be represented in negative contexts like this? These tensions would need to be dealt with in any future remediation.

In many cases, it may not be possible to simultaneously achieve all remediative ideals, especially when there is no alignment as to what bias is in the first place. This means that designers of future user-driven auditing platforms cannot fully escape making design decisions that will impact *whose values will be prioritized and how*. The design of equitable processes for recommending or deciding upon remediation strategies is a critical open problem in this space.

When fixing bias is neither desirable nor possible, promoting awareness of algorithmic biases and other harmful behaviors can be especially valuable. User awareness helps users adapt their behaviors in and around algorithmic systems, in turn fostering user-driven auditing and supporting remediation of harmful issues by other means. Participants described a wide range of ways that awareness of bias could be promoted, and most of the other remediative methods they described require user awareness in order to commence.

Beyond discussions of how algorithmic systems should be "fixed" are questions about how many outstanding issues we as a society are willing to accept and mark as 'need to be fixed' before taking more severe recourse. Other researchers have highlighted the ways that algorithmic harm derives from the structural power given them [8, 58]. Though a common reaction to harmful algorithms is to address specific issues, there have been times when the response has instead removed the algorithm's power entirely (e.g., [24, 50, 78]); indeed refusal instead of repair has been highlighted as an often appropriate way to respond to bias [11, 79]. At what point in a harmful algorithm's existence do we mitigate the harm by stopping the algorithm altogether rather than just attempting to mitigate specific issues?

7 CONCLUSION

We demonstrated everyday users' strong abilities to detect and reason about potentially harmful algorithmic behaviors through a three-stage process model (i.e., search inspiration, sensemaking, and remediation) as well as significant influences on the process (i.e., knowledge and beliefs, and platform affordances). These findings reveal many opportunities in designing user-driven algorithm audits such as utilizing users' prior exposure to and experiences of harms and biases as well as leveraging platforms' affordances for comparing and reporting potential cases of algorithmic harms. The collective effort of users in the workshops to understand, investigate, and consider remediations for potentially harmful algorithmic behaviors also showcase opportunities for building counterpublics and collective audit platforms. However, balancing the dynamics of such platforms including the demographic makeup and community guidance and feedback needs further research, along with a careful design process.

ACKNOWLEDGMENTS

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 Amazon Research Award and a Cisco Research Award. Special thanks to Jason Hong, Adam Perer, Nihar Shah, our anonymous reviewers, and all our participants.

REFERENCES

[1] [n.d.]. Google image search cements national stereotypes of 'racy' women. ([n. d.]). https://www.dw.com/en/google-image-search-cements-national-stereotypes-of-racy-women/a-56767605

[2] 2016. Do Google's 'unprofessional hair' results show it is racist? (2016). https://www.theguardian.com/technology/2016/apr/08/does-google-unprofessional-hair-results-prove-algorithms-racist-

[3] 2016. I saw a tweet saying "Google unprofessional hairstyles for work". I did. Then I checked the 'professional' ones. (2016). https://twitter.com/HereroRocher/status/717457819864272896

[4] 2018. Gender and Jobs in Online Image Searches. (2018). https://www.pewresearch.org/social-trends/2018/12/17/gender-and-jobs-in-online-image-searches/

[5] 2019. Holy $%#@: Religion & Conscious Language. panel at ACES Annual Conference. https://aceseditors.org/conference/past-conferences/2019/sessions

[6] 2019. Search Engine Bias. (May 2019). https://therepproject.org/search-engine-bias/

[7] 2020. (2020). https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/questionnaires-and-instructions/questionnaires/2020-informational-questionnaire.pdf

[8] Ali Alkhatib. 2021. *To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445740

[9] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing Race and Gender Discrimination in Online Housing Markets. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 24–35. https://ojs.aaai.org/index.php/ICWSM/article/view/7276

[10] Joshua Attenberg, Panos Ipeirotis, and Foster Provost. 2015. Beat the Machine: Challenging Humans to Find a Predictive Model's "Unknown Unknowns". *J. Data and Information Quality* 6, 1, Article 1 (mar 2015), 17 pages. https://doi.org/10.1145/2700832

[11] Eric P.S. Baumer and M. Six Silberman. 2011. When the Implication is Not to Design (Technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2271–2274. https://doi.org/10.1145/1978942.1979275

[12] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. *CoRR* abs/1801.10408 (2018). arXiv:1801.10408 http://arxiv.org/abs/1801.10408

[13] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050* (2020).

[14] Khristopher J Brooks. 2020. Twitter users say the platform crops out Black faces. *CBS News* (Sep 2020). https://www.cbsnews.com/news/twitter-image-cropping-algorithm-racial-profiling/

[15] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services. (2019), 1–12. https://doi.org/10.1145/3290605.3300271

[16] Andrew D Brown, Patrick Stacey, and Joe Nandhakumar. 2008. Making sense of sensemaking narratives. *Human relations* 61, 8 (2008), 1035–1062.

[17] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. 81 (23–24 Feb 2018), 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[18] Ángel Alexander Cabrera, Abraham Druck, Jason I. Hong, and Adam Perer. 2021. Discovering and Validating AI Errors With Crowdsourced Failure Reports. 5, CSCW2 (2021).

[19] Metz Cade. [n.d.]. 'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You? ImageNet Roulette, a digital art project and viral selfie app, exposes how biases have crept into the artificial-intelligence technologies changing our lives. ([n. d.]). https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html

[20] DJ PATIL CECILIA MUÑOZ, MEGAN SMITH. 2016. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. (2016). https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

[21] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. *Investigating the Impact of Gender on Rank in Resume Search Engines.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174225

[22] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and Addressing Algorithmic Bias in Practice. *Interactions* 25, 6 (Oct. 2018), 58–63. https://doi.org/10.1145/3278156

[23] Kate Crawford and Trevor Paglen. 2021. Excavating AI: The politics of images in machine learning training sets. *AI & Society* (2021), 1–12.

[24] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. (2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[25] MICHAEL ANN DEVITO. 2021. Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms. (2021). https://doi.org/10.1145/3476080

[26] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. *How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173694

[27] Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms Ruin Everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3163–3174. https://doi.org/10.1145/3025453.3025659

[28] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. *First I "like" It, Then I Hide It: Folk Theories of Social Feeds.* Association for Computing Machinery, New York, NY, USA, 2371–2382. https://doi.org/10.1145/2858036.2858494

[29] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. *Communicating Algorithmic Process in Online Behavioral Advertising.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174006

[30] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning about Invisible Algorithms in News Feeds. 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, 153–162. https://doi.org/10.1145/2702123.2702556

[31] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. Be careful; Things can be worse than they appear - Understanding biased algorithms and users' behavior around them in rating platforms". (2017), 62–71. Funding Information: This work was funded by NSF grant CHS-1564041. Publisher Copyright: © Copyright 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 11th International Conference on Web and Social Media, ICWSM 2017 ; Conference date: 15-05-2017 Through 18-05-2017.

[32] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. (2019), 1–14. https://doi.org/10.1145/3290605.3300724

[33] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 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.

[34] Nancy Fraser. 1990. Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social text* 25/26 (1990), 56–80.

[35] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

[36] Day S. Goldstone, R. L. and J. Y. Son. 2010. Comparison. In *Towards a theory of thinking*, Goel Vinod von Müller Albrecht Glatzeder, Britt (Ed.). Springer, Berlin, Heidelberg, 103–121.

[37] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. *The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445423

[38] Nina Grgic-Hlaca, 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. (2018), 903–912. https://doi.org/10.1145/3178876.3186138

[39] Ben Guarino. [n.d.]. Google faulted for racial bias in image search results for black teenagers. ([n. d.]). https://www.dw.com/en/google-image-search-cements-national-stereotypes-of-racy-women/a-56767605

[40] Jessica Guynn. 2015. Google photos labeled black people 'gorillas'. *USA Today* (June 2015). https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465

[41] Anikó Hannák, Piotr Sapiezynski, Arash Molavi Khaki, David Lazer, Alan Mislove, and Christo Wilson. 2017. Measuring Personalization of Web Search. *CoRR* abs/1706.05011 (2017). arXiv:1706.05011 http://arxiv.org/abs/1706.05011

[42] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring Price Discrimination and Steering on E-Commerce Web Sites. (2014), 305–318. https://doi.org/10.1145/2663716.2663744

[43] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring Price Discrimination and Steering on E-Commerce Web Sites. (2014), 305–318. https://doi.org/10.1145/2663716.2663744

[44] Alex Hern. 2020. Twitter apologises for 'racist' image-cropping algorithm. *The Guardian* (Sept. 2020). https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm

[45] Kenneth Holstein, Erik Harpstead, Rebecca Gulotta, and Jodi Forlizzi. 2020. Replay enactments: Exploring possible futures through historical data. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1607–1618.

[46] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (May 2019). https://doi.org/10.1145/3290605.3300830

[47] Samuel GB Johnson, Greeshma Rajeev-Kumar, and Frank C Keil. 2016. Sensemaking under ignorance. *Cognitive psychology* 89 (2016), 39–70.

[48] 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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 305 (oct 2021), 44 pages. https://doi.org/10.1145/3476046

[49] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. (2015), 3819–3828. https://doi.org/10.1145/2702123.2702520

[50] Anna Kramer. 2021. Twitter's image cropping was biased, so it dumped the algorithm. https://www.protocol.com/twitter-image-cropping-algorithm-biased

[51] Michael W Kramer. 2017. Sensemaking. *The international encyclopedia of organizational communication* (2017), 1–10.

[52] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. *CoRR* abs/1704.01347 (2017). arXiv:1704.01347 http://arxiv.org/abs/1704.01347

[53] Michelle Seng Ah Lee and Jatinder Singh. [n.d.]. The Landscape and Gaps in Open Source Fairness Toolkits (September 18, 2020), biburl = https://doi.org/10.1145/3411764.3445261. ([n. d.]).

[54] Rosina Lippi-Green. [n.d.]. English with an Accent: Language, Ideology and Discrimination in the United States. ([n. d.]).

[55] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. (2020), 1–14. https://doi.org/10.1145/3313831.3376445

[56] Bugcrowd Product Marketing. 2020. Reducing Noise in Crowdsourced Security. https://www.bugcrowd.com/blog/reducing-noise-in-crowdsourced-security/

[57] Aarian Marshall. 2021. Gig Workers Gather Their Own Data to Check the Algorithm's Math. https://www.wired.com/story/gig-workers-gather-data-check-algorithm-math/

[58] Milagros Miceli, Julian Posada, and Tianling Yang. 2021. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? arXiv:2109.08131 [cs.HC]

[59] Safiya Umoja Noble. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. (2018). https://www.jstor.org/stable/j.ctt1pwt9w5

[60] Rodrigo Ochigame and Katherine Ye. 2021. Search Atlas: Visualizing Divergent Search Results Across Geopolitical Borders. (2021), 1970–1983. https://doi.org/10.1145/3461778.3462032

[61] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[62] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *CoRR* abs/2001.00973 (2020). arXiv:2001.00973 http://arxiv.org/abs/2001.00973

[63] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 148 (Nov. 2018), 22 pages. https://doi.org/10.1145/3274417

[64] Aja Romano. 2019. A group of YouTubers is trying to prove the site systematically demonetizes queer content. (2019). https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report

[65] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. (1993), 269–276.

[66] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* (2014).

[67] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 154 (nov 2018), 19 pages. https://doi.org/10.1145/3274423

[68] Sarita Schoenebeck. [n.d.]. ([n. d.]). https://twitter.com/syardi/status/1276587095767887874?s=20

[69] Nick Seaver. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4, 2 (2017), 2053951717738104. https://doi.org/10.1177/2053951717738104 arXiv:https://doi.org/10.1177/2053951717738104

[70] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. (2019), 59–68. https://doi.org/10.1145/3287560.3287598

[71] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 433 (oct 2021), 29 pages. https://doi.org/10.1145/3479577

[72] Ignacio Siles, Andrés Segura-Castillo, Ricardo Solís, and Mónica Sancho. 2020. Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global South. *Big Data & Society* 7, 1 (2020), 2053951720923377. https://doi.org/10.1177/2053951720923377 arXiv:https://doi.org/10.1177/2053951720923377

[73] Jina Suh, Soroush Ghorashi, Gonzalo Ramos, Nan-Chen Chen, Steven Drucker, Johan Verwey, and Patrice Simard. 2019. AnchorViz: Facilitating Semantic Data Exploration and Concept Discovery for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 10, 1, Article 7 (Aug. 2019), 38 pages. https://doi.org/10.1145/3241379

[74] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. (2013). arXiv:1301.6822 [cs.IR]

[75] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M. Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 28 (Nov. 2019), 21 pages. https://doi.org/10.1145/3359130

[76] Julia Velkova and Anne Kaun. 2019. Algorithmic resistance: Media practices and the politics of repair. *Information, Communication & Society* (2019), 1–18.

[77] Neil Vigdor. 2019. Apple card investigated after gender discrimination complaints. The New York Times. (2019). https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

[78] James Vincent. 2016. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge* (Mar 2016). https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

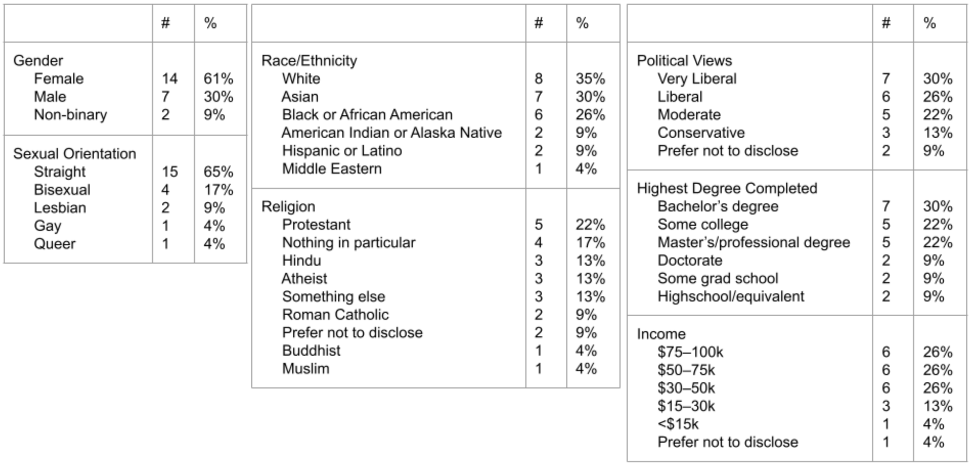
[79] Cedric Deslandes Whitney, Teresa Naval, Elizabeth Quepons, Simrandeep Singh, Steven R Rick, and Lilly Irani. 2021. *HCI Tactics for Politics from Below: Meeting the Challenges of Smart Cities.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445314

[80] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. (2018), 1–14. https://doi.org/10.1145/3173574.3174230

[81] Meg Young, Lassana Magassa, and Batya Friedman. [n.d.]. Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology* ([n. d.]).

[82] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist—it's time to make it fair. *Nature* 559 (2018), 324–326.

A ADDITIONAL DEMOGRAPHIC INFORMATION



	#	%
Gender		
Female	14	61%
Male	7	30%
Non-binary	2	9%
Sexual Orientation		
Straight	15	65%
Bisexual	4	17%
Lesbian	2	9%
Gay	1	4%
Queer	1	4%

	#	%
Race/Ethnicity		
White	8	35%
Asian	7	30%
Black or African American	6	26%
American Indian or Alaska Native	2	9%
Hispanic or Latino	2	9%
Middle Eastern	1	4%
Religion		
Protestant	5	22%
Nothing in particular	4	17%
Hindu	3	13%
Atheist	3	13%
Something else	3	13%
Roman Catholic	2	9%
Prefer not to disclose	2	9%
Buddhist	1	4%
Muslim	1	4%

	#	%
Political Views		
Very Liberal	7	30%
Liberal	6	26%
Moderate	5	22%
Conservative	3	13%
Prefer not to disclose	2	9%
Highest Degree Completed		
Bachelor's degree	7	30%
Some college	5	22%
Master's/professional degree	5	22%
Doctorate	2	9%
Some grad school	2	9%
Highschool/equivalent	2	9%
Income		
$75–100k	6	26%
$50–75k	6	26%
$30–50k	6	26%
$15–30k	3	13%
<$15k	1	4%
Prefer not to disclose	1	4%

Figure 3: self-reported participant demographic information from the screening questionnaire.

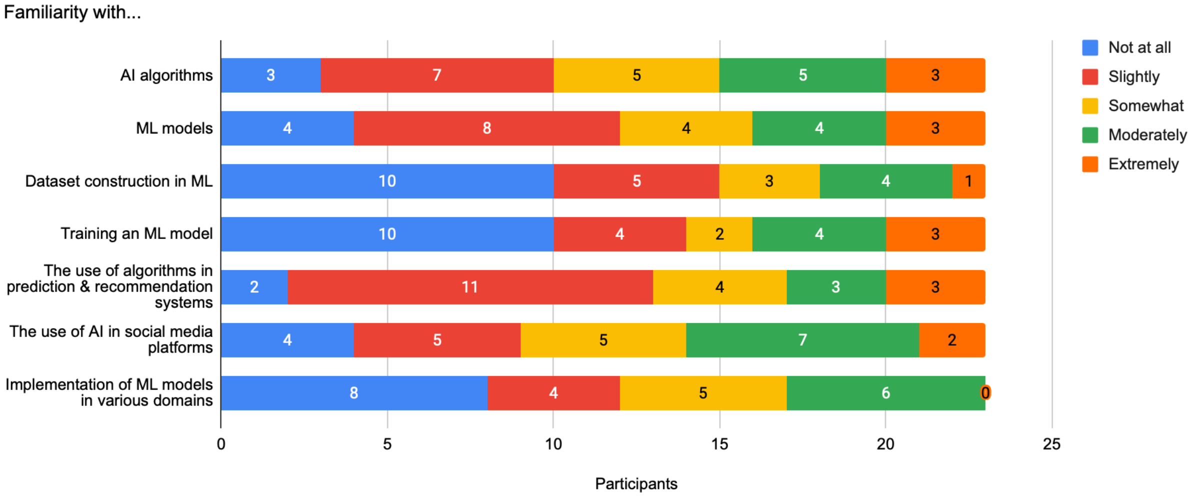


Figure 4: participant familiarities with various algorithmic concepts, on a 5-point Likert scale.