Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice

Wesley Hanwen Deng
hanwend@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Bill Boyuan Guo
boyuang@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Alicia DeVrio
adevos@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

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

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

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

ABSTRACT

Recent years have seen growing interest among both researchers and practitioners in *user-engaged* approaches to algorithm auditing, which directly engage users in detecting problematic behaviors in algorithmic systems. However, we know little about industry practitioners' current practices and challenges around user-engaged auditing, nor what opportunities exist for them to better leverage such approaches in practice. To investigate, we conducted a series of interviews and iterative co-design activities with practitioners who employ user-engaged auditing approaches in their work. Our findings reveal several challenges practitioners face in appropriately recruiting and incentivizing user auditors, scaffolding user audits, and deriving actionable insights from user-engaged audit reports. Furthermore, practitioners shared organizational obstacles to user-engaged auditing, surfacing a complex relationship between practitioners and user auditors. Based on these findings, we discuss opportunities for future HCI research to help realize the potential (and mitigate risks) of user-engaged auditing in industry practice.

KEYWORDS

user-engaged algorithm auditing, responsible AI, industry practitioners, fairness, bias

ACM Reference Format:
Wesley Hanwen Deng, Bill Boyuan Guo, Alicia DeVrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *CHI '23: ACM Conference on Human Factors in Computing Systems, April 23–38, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '23, April 23–28, 2023, Hamburg, Germany

© 2023 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years, algorithm audits have risen to prominence as an approach to uncover biased, discriminatory, or otherwise harmful behaviors in algorithmic systems [16, 22, 29, 31, 46, 60, 68, 77, 86, 95]. Today, algorithm audits are typically conducted by small groups of experts such as industry practitioners, researchers, and activists [60, 77]. Although expert-led approaches have been highly impactful, they often suffer from major blindspots, and fail to detect critical issues. For example, expert-led audits can fail when those conducting the audit lack the relevant cultural knowledge and lived experience to recognize and know where to look for certain kinds of harmful algorithmic behaviors [25, 42, 80, 92].

To overcome limitations of current algorithm auditing techniques, researchers in HCI and AI have begun to explore the potential of more *user-engaged* approaches to algorithm auditing, which directly engage users of AI products and services in surfacing harmful algorithmic behaviors. Recent years have seen many cases in which users organically came together to uncover and raise awareness about harmful behaviors in algorithmic systems they use day-to-day, which had eluded detection by industry teams or other expert auditors [80]. Inspired by these observations, researchers have begun to explore the design of systems that can leverage the power of everyday users and crowds to surface harmful algorithmic behaviors that might otherwise go undetected (e.g., [8, 17, 48, 53, 63, 64]). The designs of existing research systems span a spectrum of user engagement, from more practitioner-led approaches—such as crowdsourcing workflows in which users' testing and auditing activities are more heavily guided and constrained by requesters—to more user-led approaches in which users take greater initiative in directing their own activities.

In parallel to these research efforts, several major technology companies have begun to experiment with approaches that engage users in auditing their AI products and services for problematic behaviors. For example, in 2021 Twitter introduced its first "algorithmic bias bounty" challenge to engage users in identifying harmful biases in its image cropping algorithm [21]. In another effort, Google launched the "AI Test Kitchen," a web-based application that invites users to experiment with some of Google's latest AI-based conversational agents, and to report any problematic behaviors they encounter [90]. More recently, inspired by Twitter's "algorithmic bias bounty," OpenAI initiated a "Feedback Contest"

to encourage users to "provide feedback on problematic model outputs" during their interactions with ChatGPT chatbot [4].

Despite growing interest from industry, there remains a gulf between the academic research literature on user-engaged auditing and current industry practice. In particular, we still know little about industry AI practitioners' current practices and challenges around user-engaged auditing, and what opportunities exist for them to better leverage such approaches in practice. To investigate, in this paper we explore the following research questions:

RQ1 What are AI practitioners' current motivations and practices around user engagement in auditing their AI products and services for problematic algorithmic behaviors?

RQ2 What opportunities and challenges do practitioners envision for user-engaged approaches to better support their algorithm auditing efforts?

We conducted a two-stage study with 12 industry practitioners from 9 technology companies, all of whom have experimented with engaging users in auditing their AI systems for problematic algorithmic behaviors. We first conducted semi-structured interviews to understand practitioners' current practices and challenges around engaging users in AI testing and auditing. We then conducted co-design activities, working with practitioners to iteratively co-design three design artifacts as a way to further probe challenges perceived by practitioners and opportunities to better support user-engaged approaches to algorithm auditing in industry practice.

Overall, our participants shared three major motivations for engaging users in AI testing and auditing: understanding users' subjective experiences of problematic machine behaviors, overcoming their teams' blindspots when auditing their products and services, and gathering evidence from users to help them advocate for fairness work within their organizations. Participants shared prior experiences engaging users on different scales, from individual user study sessions, to focus group workshops, to large-scale user feedback and crowdsourcing activities. However, in doing so, practitioners encountered various challenges in engaging users effectively. For instance, practitioners discussed challenges they faced in recruiting and incentivizing the "right" group of auditors for a given task, with relevant identities and lived experiences. Participants also discussed the difficulties in scaffolding users towards productive auditing strategies, without biasing them to simply replicate industry teams' own blindspots. Finally, practitioners discussed the challenges of quantification when deriving actionable insights from user-engaged auditing reports: relying upon the majority vote runs the risk of masking the very biases an audit is intended to uncover. In addition, participants shared broader organizational obstacles to user-engaged auditing, highlighting key tensions that arise in practice when involving users in algorithm auditing efforts such as potential PR risks, profit motives that work against protecting marginalized groups, and privacy and legal concerns.

As private companies increasingly experiment with user-engaged approaches to algorithm auditing, HCI research has a critical role to play in shaping more effective and responsible practices. To this end, this work contributes:

- An in-depth understanding of industry practitioners' motivations, current practices, and challenges in effectively engaging users in testing and auditing AI products and services. Our findings shed light on the types of problems practitioners aim to address through user engagement around algorithm auditing, as well as the the ways practitioners navigate organizational tensions around user involvement in AI development processes.
- A set of design implications for user-engaged algorithm auditing, beyond standard considerations for human computation or user feedback systems.
- Insights into the complex relationship between user auditors and industry practitioners working on responsible AI, suggesting opportunities for future HCI research to help realize the potential (and mitigate risks) of user-engaged auditing in industry practice.

2 RELATED WORK

2.1 Understanding and supporting responsible AI practices in industry contexts

In recent years, significant effort has been directed towards the development of approaches, guidelines, and tools to help industry practitioners audit their AI products and services for unfair, biased, or otherwise harmful algorithmic behaviors (e.g., [3, 11, 13, 14, 61, 71, 72]). Early work in this area has largely been guided by advances in academic research on AI fairness [1, 6, 10, 32, 52, 67]. Yet in a series of interview studies and surveys with industry AI practitioners, Holstein et al. [42] found that there were major disconnects between the tools offered by the research community, versus the actual on-the-ground needs of industry AI practitioners. To address such gaps, a growing line of research in HCI has focused on better understanding industry AI practitioners needs and designing to support responsible AI practices in industry. For example, studies from Madaio et al. [56] and Rakova et al. [70] investigated the organizational challenges and barriers that practitioners face in practice when attempting to build more responsible AI systems.

Meanwhile, to better support responsible AI practices, companies have been developing responsible AI guidelines such as People + AI guidebook [72], trustworthy AI principles [87], AI fairness checklists [71], and responsible AI toolkits such as AI Explainability 360 [3] and Fairlearn [13]. However, recent HCI research has surfaced gaps between fairness toolkits' capabilities and practitioners' needs [24, 54, 54, 73]. For example, Kaur et al. [47] found that AI practitioners often over-trust and misuse AI explainability toolkits. Other work from Lee et al. and Deng et al. identified misalignment between the designs of existing fairness toolkits versus practitioners' actual desires and usage of these toolkits [24, 54, 73]. In interviews with AI practitioners, these authors found that, beyond the functionality provided by current toolkits, practitioners desired tools that could help them bring in perspectives from relevant domain experts and users, in order to aid them in auditing their AI systems [24]. In the next sections, we discuss emerging work that aims to harness the power of users in algorithm auditing.

2.2 The power of users in algorithm auditing

Metaxa et al. [60] define an algorithm audit as "a method of repeatedly querying an algorithm and observing its output to draw

conclusions about the algorithm's opaque inner workings and possible external impact." A growing body of work in HCI, AI, and related communities has developed tools and processes to audit algorithmic systems for biased, discriminatory, or otherwise harmful behaviors (e.g., [16, 60, 77]). Past work in algorithm auditing has uncovered harmful algorithmic behaviors across a wide range of algorithmic systems, from search engines to hiring algorithms to computer vision applications [7, 16, 37, 62, 66, 86].

Today, algorithm audits are typically conducted by small groups of experts such as industry practitioners, researchers, activists, and government agencies [60]. However, such expert-led audits often fail to surface serious issues that everyday users of algorithmic systems are quickly able to detect once a system is deployed in the real world [42, 80]. For instance, this approach can fail when those conducting the audit lack the relevant cultural knowledge and lived experience to recognize and know where to look for certain kinds of harmful algorithmic behaviors [25, 42, 80, 92]. In addition, expert-led audits may fail to detect certain harmful algorithmic behaviors because these behaviors only arise—or are only recognized as harmful—when a system is used in particular context or in particular ways, which auditors may fail to anticipate [22, 30, 34, 42, 79, 80].

Recent years have seen many real-world cases in which *users* have uncovered and raised awareness around harmful algorithmic behaviors in systems they use day-to-day (e.g., search engines [16], online rating/review systems [29, 86], and machine translation systems [66]) although expert auditors had failed to detect these issues. Shen et al. [80] developed the concept of "everyday algorithm auditing" to describe how everyday users detect, understand, and interrogate problematic machine behaviors via their daily interactions with algorithmic systems. In the cases these authors reviewed, regular users of a wide range of algorithmic systems and platforms came together organically to hypothesize and test for potential biases. More recently, DeVos et al. [25] conducted a series of behavioral studies to better understand how users are often able to be so effective, both individually and collectively, in surfacing harmful algorithmic behaviors that more formal or expert-led auditing approaches fail to detect. As discussed next, recent research is beginning to explore ways to harness the power users in algorithm auditing to overcome limitations of expert-led approaches.

2.3 Supporting user-engaged algorithm auditing

Recognizing the power of users in algorithm auditing, researchers have begun to explore the design of systems to support more *user-engaged* approaches [25, 53] to algorithm auditing, which directly engage users in surfacing harmful algorithmic behaviors that might otherwise go undetected.

A line of work has developed interfaces, interactive visualizations, and crowdsourcing pipelines to support people in actively searching for algorithmic biases and harmful behaviors [8, 17, 48, 63]. The designs of these research systems span a spectrum of user-engagement, from more practitioner-led approaches to more user-led approaches in which users take greater initiative and control in directing their efforts. For example, Ochigame and Ye developed a web-based tool called Search Atlas, which enables users to easily conduct side-by-side comparisons of the Google search results they might see if they were located in different countries to spot [64]. Kiela et al. developed a general research platform called Dynabench, which invites users to try to identify erroneous and potentially harmful behaviors in AI models [48]. Using Dynabench, users can generate test inputs to a model to try to find problematic behaviors, flag behaviors they identify, and provide brief open-text responses if they wish to offer additional context. More recently, Lam et al. developed a tool called "IndieLabel," in order to empower end users to detect and flag potential algorithmic biases and then author audit reports to communicate these to relevant decision-makers [53].

In parallel, several major technology companies have begun to experiment with approaches that engage users in auditing their AI products and services for problematic behaviors. For example, in 2021 Twitter introduced its first "algorithmic bias bounty" challenge to engage users in identifying harmful biases in its image cropping algorithm [21]. In another effort, Meta adopted the Dynabench platform described above, to discover potentially harmful behaviors in natural language processing models [48]. More recently, Google launched the "AI Test Kitchen," a web-based application that invites users to experiment with Google's latest LLMs-powered conversational agents, and to report any problematic behaviors they encounter, with the stated goal of engaging users in *"learning, improving, and innovating responsibly on AI together"* [90]. In addition, organizations like OpenAI and HuggingFace are beginning to include built-in interface features that invite users to report harmful algorithmic behaviors they encounter while interacting with LLM-powered applications like text-to-image generation tools. HuggingFace developed features to engage end users in flagging ethical/legal issues on their API [65]. In addition, OpenAI initiated a feedback contest around their LLM-based tool ChatGPT, with the goal of encouraging users to "provide feedback on problematic model outputs" [4].

Despite growing interest in both academia and industry, there remains a gulf between the academic research literature on user-engaged auditing and current industry practice. To date, little is known about industry AI practitioners' current practices and challenges around user-engaged auditing, nor what opportunities exist for them to better leverage such approaches in practice. In this paper, we take a first step towards understanding current practices, challenges, and design opportunities for user-engaged approaches to algorithm auditing in industry practice.

3 METHOD

3.1 Study design

We conducted a two-stage study involving semi-structured interviews followed by iterative co-design activities. We first conducted semi-structured interviews to understand participants' current practices and challenges around engaging users in AI testing and auditing. In the next stage, we engaged participants in a co-design activity to further probe the opportunities and challenges in supporting user-engaged algorithm auditing in industry practice. We worked with participants to iteratively design three artifacts: a **user-engaged audit report**, representing a "wish list" of types of information that they would ideally want to solicit through a user-engaged auditing approach, and **two user-engaged auditing pipelines**, building upon initial designs informed by interview findings and insights from prior literature [25, 60, 77, 80]. Throughout

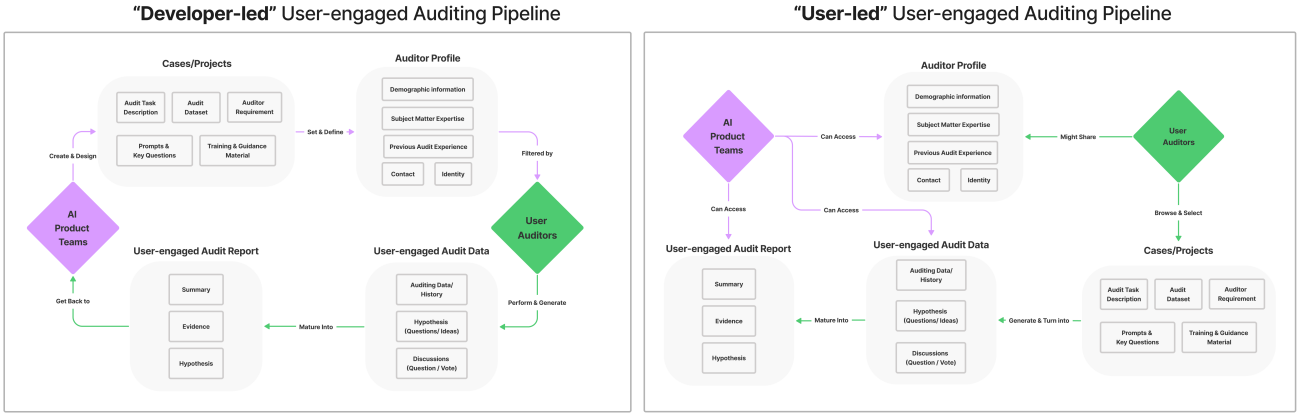

Figure 1: Two potential user-engaged auditing pipeline designs that were iteratively co-designed with participants. The left image shows the "developer-led" pipeline design, and the right image shows the "user-led" pipeline design. Each figure illustrates a possible interaction flow between user auditors and AI product teams, showing how auditing tasks are created, how background information on user auditors is shared, how user auditing reports are generated based on auditors' findings, and how these reports are shared with AI product teams. During the co-design activity, participants could zoom in, annotate, and modify the details. We used these pipeline flowcharts as probes, not as final products, to investigate more deeply on practitioners' challenges and desires.

the study, we iterated on these design artifacts based on feedback and design ideas from prior participants. We used these artifacts and the *process* of co-designing them to provoke deeper conversations around participants' desires, as well as potential risks they anticipate, for new systems that support user-engaged auditing.

Following an iterative co-design process similar to prior work (cf. [41, 57]), for our first five participants, we ran the two stages of our study in separate sessions in order to better inform the design of the initial versions of the artifacts based on the needs and desires these participants expressed in the first set of interviews. However, we soon encountered difficulty in retaining industry participants due to their busy schedules (e.g., one participant was not able to return to complete the co-design). Therefore, after our first five participants, we ran both stages in a single session. We then continued to iterate on the artifacts during the study sessions themselves. Below, we describe each of these activities in more detail.[1]

3.1.1 Stage one: Semi-structured interviews. To understand practitioners' current practices around engaging users in AI testing and auditing, we conducted semi-structured interviews, each lasting up to an hour. We adopted a directed storytelling approach [33]. We first asked participants to describe their team's prior experiences in trying to detect or address biased or harmful behaviors in their AI products or services, with a specific focus on whether, why, and how they engaged users in the process. For example, we asked *"Could you describe how your team attempted to engage users in auditing the AI products and services you mentioned"* and *"What motivated you or your team to engage users in this way?"* Through follow-up questions, we probed deeper into challenges participants had encountered when attempting to engage users in the auditing process. As participants shared specific challenges they had encountered, we also invited them to share ideas for potential solutions to these challenges. For example, in response to specific challenges raised by participants, we asked *"How did your team attempt to tackle these challenges?"* and *"How effective were your team's approaches?"*

3.1.2 Stage two: Iterative co-design activities. To further envision future opportunities and solicit potential challenges and risks for user-engaged algorithm auditing approaches, following the interviews, we then involved participants in a series of co-design activities, following an iterative co-design process similar to prior work (cf. [41, 57]). This stage of the study lasted up to 45 minutes, and involved participants in co-design around three design artifacts: a **user-engaged audit report** and two **user-engaged audit pipeline flowcharts**. We first designed initial versions of these artifacts based on participant needs and desires expressed during stage one, as well as prior research on user-engaged algorithm auditing [25, 80]. We then iterated on their designs with practitioners throughout the co-design activities. We note that these design artifacts were not the goal of our study, but rather served as tools to probe more deeply on practitioners' challenges and desires. Below, we describe the process of co-designing these three artifacts, and how we used this process to probe on future opportunities and risks of user-engaged audits.

User-engaged audit report: We invited each participant to contribute to the design of a report that they would ideally like to see as the *output* of a user-engaged auditing process. We first asked participants open-ended questions such as *"What information would you ideally want the service to report back to your team?"* and encouraged them to sketch as they generated new ideas. To help participants come up with ideas, we presented participants with example of actual written responses generated by users during a

[1]We also provide our interview and co-design protocol in the supplementary material.

user-engaged algorithm auditing workshop in prior work[2] [25]. Viewing examples of actual user responses on an auditing task provided an opportunity for participants to reflect upon gaps in their own report designs, and to anticipate potential challenges in soliciting useful information from user auditors (cf. [40, 57]). We then presented participants with a user-engaged audit report template (see Fig. 3 in Appendix) that we initially designed based on previous work and iterated through the design during the study. Our goal was to further probe participants' feedback by providing them with potential content that an audit report can consist of (such as information about the auditors, details of the reported issue, evidence, the severity of the issue, etc.). However, we intentionally showed this report after participants generated ideas about what a report can include to avoid biasing them towards a specific report format, yet giving them the opportunity to discuss other options and iterating through the report design based on their initial ideas.

User-engaged audit pipeline flowcharts: We also co-designed two opposing caricatures of user-engaged audit pipeline designs with participants which varied in the degree of initiative users assumed in the auditing process: (1) a **"developer-led"** pipeline, in which the audits were primarily initiated and coordinated by the developers of an AI product or service; (2) a **"user-led"** pipeline, in which the audits were primarily initiated and coordinated by users (see Fig. 1). For the developer-led pipeline, the AI product teams can fully control what AI systems (or what parts of the AI systems) should be audited, who should be considered eligible auditors, and how the auditing should be executed. In the "user-led" pipeline, users are those who initiate and control the auditing based on their interactions with the AI systems. Users could collectively initiate the auditor selection criteria, defining auditing protocol, generating audit data, and synthesizing reports. In this case, AI practitioners can only access the audit data without getting involved in or having a say in the auditing process. While we anticipated that *neither* of these caricatures would represent ideal designs from industry practitioners' perspectives, we presented these in order to *provoke further discussion* about design trade-offs between greater user versus developer control in auditing processes.

3.2 Participants

We adopted a purposive sampling approach [19], with the aim of recruiting industry practitioners who either (1) had direct prior experience employing *user-engaged algorithm auditing* approaches, or (2) had an interest in such approaches and had adjacent experience *crowdsourcing* approaches as part of their AI work. Specifically, using an online screening survey, we recruited members of industry teams that design and build AI products and services, and who had *already* attempted to engage users in detecting fairness-related issues in their AI systems. In addition, we opened up the study to interested practitioners who had *not* yet experimented with user-engaged approaches to algorithm auditing, but who had prior experience using crowdsourcing approaches in other areas of their AI work. We broadened our criteria to include these participants because we expected that prior experience with crowdsourcing would help participants envision ways user-engaged approaches might support their algorithm auditing efforts. In the end, however, *all* 12 of our participants had direct prior experience experimenting with user-engaged approaches to algorithm auditing (see Table 2). In addition, all but three of our participants had prior experience with crowdsourcing methods.

We recruited our participants through social media (e.g., Linkedin and Twitter), and through direct contacts at large technology companies. As discussed in prior literature that studies responsible AI practices in industry (e.g., [24, 42, 54]), recruitment for such studies can be highly challenging. Practitioners are often wary of participating in such interview studies, for instance, given that participation may require admitting the existence of flaws in their products and services that have not been made public, or sharing disagreements about their companies' current organizational culture. Although we assured potential participants that their responses would be carefully de-identified, as discussed below, we expect that such concerns likely had an influence on our recruitment.

In total, 25 practitioners completed the recruitment screening form, of which 18 met our recruitment criteria. Ultimately, 12 of these practitioners, spanning 9 companies, responded to our study invitation and participated in the study. All 12 participants participated in the interview session; all except one participant participated in the co-design session. (P4 was not able to return to complete the co-design activity due to busy schedule.) All participants were compensated at a rate of $35 per hour for their participation. Table 1 overviews participants' job titles, their years of experience with user-engaged auditing, their company size, the types of AI products or services they worked on, and their experiences with user-engaged auditing. While 9 of our participants conducted user-engaged algorithm auditing as part of their main job function; three participants (P2, P6, P12) engaged users in algorithm auditing on their own initiative, to help them in advocating for fairness issues to be addressed within their organizations.

Following prior work on responsible AI practices in industry [24, 42, 56, 57], to avoid identifying individual participants who work at the forefront of sensitive topics, details about participants' demographics are omitted, and we abstract some details about participants' companies and roles. In addition, we assured participants that we would not ask them to reveal any confidential or personally identifying information about their colleagues and that we would de-identify all responses at the individual, team, and organization levels. Finally, participants were instructed that they were free to skip any questions they were uncomfortable answering, or to leave the session at any time for any reason.

3.3 Data analysis

Our study sessions yielded approximately 15 hours of audio that we transcribed. To analyze our interview and co-design session transcripts, we adopted a reflexive thematic analysis approach [15]. Two of the authors met after each interview and co-design session to conduct an interpretation session, and then conducted open coding of the transcripts. Throughout this coding process, the authors continuously discussed discrepancies in interpretation, and iteratively refined the codes based on these discussions [15, 59]. In total, we generated around 1,125 unique codes. Through an iterative, bottom-up affinity diagramming process, we grouped codes

[2]We provide the example user report we used in the study in the supplementary material

	Company size	Job title	Types of AI products and services	Experience with user-engaged auditing	Years of AI fairness experience	Is fairness work part of their official role?
P1	1000–4,999	Director and Product Lead	AI-powered knowledge graph for academic literature search	Engaging search users in assessing potential biases in the underlying knowledge graph.	4	Yes
P2	10–50	Senior Director	ML model to predict potential donors	Engaging marginalized community members in surfacing potential biases in their team's ML model	3	No (self motivated)
P3	25,000+	ML Engineer	Natural language processing (NLP) applications	Engaging users in rating the risk of representational harms	2	Yes
P4	25,000+	Senior Technical Lead	A diverse range of AI products built by their customers	Engaging users in auditing a range of AI products built by their customers	3	Yes
P5	25,000+	Senior Product Manager	Sentiment analysis; OCR recognition	Leading several groups of AI product teams on engaging users in testing and auditing their AI products	5	Yes
P6	25,000+	UX Researcher	NLP-powered products (e.g., conversational agents)	Engaging users in testing conversational agents for potentially harmful behaviors	2	No (self motivated)
P7	25,000+	UX Researcher	Computer vision applications (e.g., image search)	Building an interface to engage users in auditing their image search engine	2	Yes
P8	25,000+	ML Researcher	Information retrieval and image processing	Building internal crowdsourcing tools for AI auditing	3	Yes
P9	5,000 - 24,999	Data Scientist	Recommendation system	Engaging users and impacted stakeholders in auditing their recommendation system	3	Yes
P10	25,000+	UX Researcher	Large language model-based conversational agent and generative image model	Building a web-based application to engage users in flagging the potential biased and harmful behavior in a conversational agent	2	Yes
P11	5,000 - 24,999	Senior Researcher	Recommendation system	Leading research efforts and producing concrete organizational policy around user-engaged algorithm auditing	2	Yes
P12	25,000+	UX Researcher	NLP	Engaging marginalized communities in auditing biases in their NLP products for low-resource languages	3	No (self motivated)

Table 1: Summary of participants' backgrounds and relevant experience.

into successively higher-level themes. The first level clustered our 1,125 codes into 271 themes. These were then clustered into 59 second-level themes, 15 third-level themes, and three final themes. We present our results in the following section, organized around our three final top-level themes.

4 FINDINGS

In this section, we describe how industry practitioners navigate the complicated process of engaging users in surfacing harmful algorithmic behaviors. As discussed below, practitioners mediate conflicts between (1) the underlying values of user engagement in algorithm testing and auditing, (2) inherent challenges in effective user engagement, and (3) the cultural, legal, and organizational obstacles that disincentivize bringing users' voices into algorithm auditing processes. We show that practitioners see clear advantages of engaging users in and around algorithm auditing processes, which have led them to explore leveraging crowdsourcing platforms and the design of new interfaces that enable in-situ feedback from their users. Yet practitioners also discussed complexities of engaging users in surfacing algorithmic harms, which introduce unique challenges beyond those faced with conventional human computation or user feedback systems. We summarize our findings in Figure 2.

4.1 Practitioners' motivations and practices for engaging users to audit AI products and services

Participants believed that engaging users in testing and auditing AI products and services could help them understand users' subjective experiences of problematic machine behaviors and help to overcome developer teams' blindspots. Participants also noted that having direct reports of potential issues from users could serve as powerful ammunition when making the case internally for a particular course of action. Driven by these motivations, participants reported having

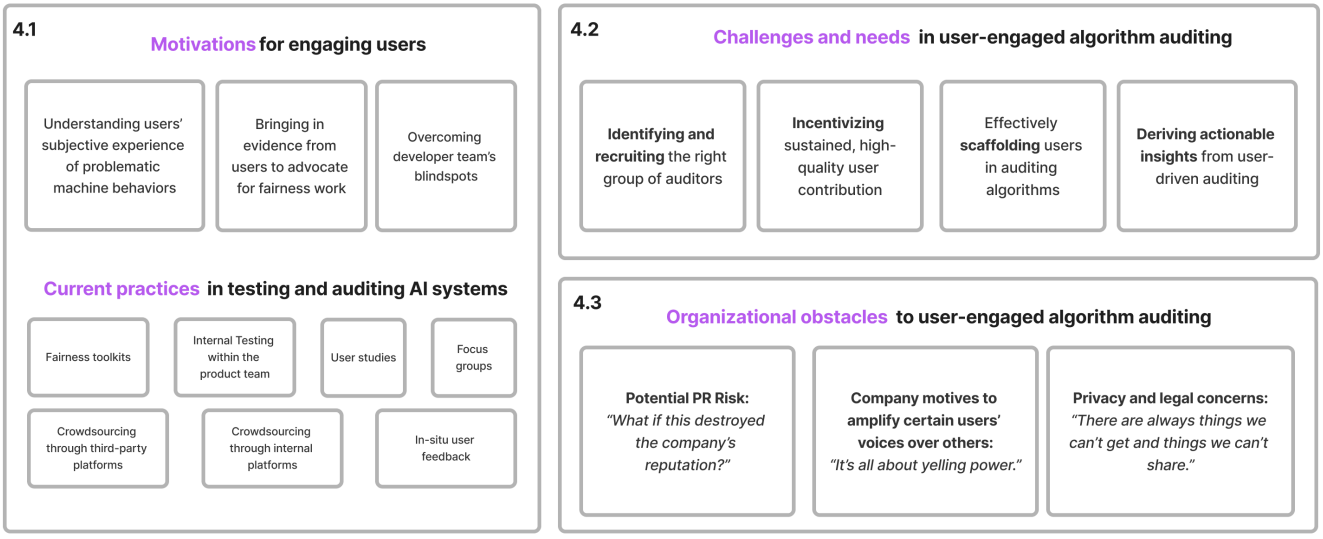

Figure 2: High-level overview of our findings. We first share participants' existing motivations and practices for user engagement around algorithm testing and auditing (Section 4.1). We then describe challenges participants have faced in their attempts to effectively engage users (Section 4.2). Finally, we share broader organizational obstacles to user-engagement around algorithm auditing perceived by participants, highlighting key tensions that arise in practice when involving users in algorithm auditing efforts (Section 4.3).

experimented with user engagement at various scales to test and audit their systems for problematic behaviors.

4.1.1 ***Motivations to engage users in auditing AI products and services.*** In our semi-structured interviews, all 12 participants reported that they currently use fairness toolkits or internal auditing with the developer teams to detect problematic behaviors in their AI products. However, participants quickly encountered limitations with these approaches, which prompted them to explore user-engaged approaches as an alternative. Below, we discuss three major motivations that participants shared for experimenting with more user-engaged approaches to algorithm auditing.

Understanding users' subjective experiences of problematic machine behaviors: Participants were motivated to adopt user-engaged approaches because in many settings, they felt that it was not possible to measure "unfairness" or "harmfulness" without an understanding of users' perceptions. As shown in Table 2, eight of our 12 participants had previously used open-source ML fairness toolkits (e.g., Fairlearn [13], AIF360 [2]) to attempt to audit their teams' AI systems. Participants said these toolkits only offered aggregate fairness metrics, whereas appropriately assessing algorithmic biases and harms in their systems required in-depth, qualitative assessments (e.g., for representational harms [23, 42]). For example, P3 shared that their main role initially involved quantitatively assessing potential biases in NLP applications using fairness toolkits. However, P3's team soon sought out user voices, because *"things like representational harms were difficult to quantify and analyze using fairness toolkits, without understanding the reason behind why users with certain identit[ies] feel uncomfortable or offended."* To this end, P3's team conducted focus groups with users who encountered harmful biases in their NLP applications, in order to better understand users' *experiences* in their own words.

Several participants described sometimes needing help from users to *determine* what "fair" or "non-harmful" behavior would look like in the first place. For example, P9 said a common approach when auditing a recommendation system is to assess *"how diverse the recommendations on your platform are."* However, different users often have conflicting perspectives regarding what it means to have appropriately "diverse" recommendations. P9 believed the way to understand this was to *"directly ask them[users] what they think."* Similarly, P11 emphasized that in order to empower their team to design effective remediations, *"It's not enough to just have people flag that a recommendation is stereotypical, we want to understand why they think it is so that our team could brainstorm potential solutions."*

Overcoming developer teams' blindspots: Another motivation for adopting user-engaged testing and auditing approaches was to overcome cultural and experiential blindspots among product team members. Ten of our 12 participants said they conducted internal testing within the product team (see Figure 2, column 2). However, echoing findings from Holstein et al. [42], participants reported that internal testing with small groups of developers often resulted in blindspots, prompting them to involve users to surface *"unknown unknowns"* (P10), or the issues that the team does not know exist. For example, P8 noted that the images on their platform came from all over the world, involving signs and languages that were deeply cultural and regional, so *"it would never be possible for [their] team to capture all these diverse aspects."* Similarly, P5 complained that "a lot of [their] AI ethics and bias activities only contain [their] own employees, and the perspectives are extremely

limited." P5 found that bringing in perspectives from external stakeholders helped to surface problematic algorithmic behaviors that their internal teams had never considered.

Some participants specifically emphasized the importance of engaging marginalized community members in auditing their AI products. P12 said, *"No one in the developing team speaks the language and knows the idioms — how would they properly audit the outcomes? That's why I have been spending time bringing native speakers into the auditing process."* When P11's team first started auditing the recommendation system in their product, they attempted to use personas to simulate real-world stakeholders and bring empathy to the team members, but they soon realized their blindspots persisted, as *"the hypothetical personas are still basically what we imagined based on our understanding [...] it's just not realistic for a group of white people to truly understand Black artists' perspectives through fictional cards."*

Bringing in evidence from users to advocate for fairness work: Finally, five participants shared that one of their major motivations for adopting user-engaged approaches was to gather evidence, in the form of direct quotes from users, that could help them persuade others on their teams that an issue was worth addressing. This was a particularly important motivation among three participants (P2, P6, P12) who were *self-motivated* to address fairness issues in their teams' products and not in roles that directly incentivized and supported this work. For example, P2 shared that, while building AI services to predict potential donors to their client institution, conversations with donors from marginalized communities inspired their team to *"fundamentally reevaluate the potential biases towards who will donate in their dataset."* As another example, P6 said their *"engineers were pretty confident about the performance of [their] model and just making assumptions about the real-world situation when people are using the tool actively,"* rarely communicating with UX researchers like P6. However, P6 *"got the luxury to chat directly with folks who are working on building the product"* when they brought in users' voices and activities about the tool. Similarly, P4 shared that when data scientists voice concerns about the ML model being unfair and biased using numbers and graphs, *"it is not a tangible risk to a business owner [...] bad experiences and feedback from the users work way better than numbers to motivate business owners to think deeper about the product's potential negative impact."*

4.1.2 *Existing algorithm auditing approaches that engage with users.*

Driven by the motivations discussed above, all 12 of our participants reported having experimented with user-engaged approaches to test and audit their systems for problematic algorithmic behaviors. As shown in Table 2, these methods varied in scale: practitioners conducted user studies and focus groups with small groups of users, but they also attempted to engage a larger number of users in the auditing process through crowdsourcing tools or in-situ user feedback.

User studies and focus groups: Nine participants shared that they have conducted single-person user studies, and eight said they have conducted focus groups, to engage users in testing their AI services and products. Participants shared how working with users provided them with different and new perspectives on a product and its potential biases. To expand their perspectives, some practitioners specifically sought out users who belonged to marginalized groups. For example, in order to mitigate potential biases in their donor database—which was used as the training data set for their AI service to identify potential future donors—P2's team connected with prior donors who have minority backgrounds and were *"not being included in the current database [that was] full of rich white people"* in order to understand their donating perspectives and experiences. This approach helped P2's team realize that their database prioritizes privileged race and gender groups and change their overall problem formulation and development strategy for their models. Similarly, P12 helped their machine translation team better understand how errors in their products impacted marginalized immigrant communities through *"chatting with immigrants living in the US with low language proficiency both individually and in groups."*

Crowdsourcing and in-situ user feedback approaches: Although talking to users provided practitioners with new perspectives on potential harms their products might introduce, practitioners desired more scalable approaches to reach larger and more diverse groups. As a result, 9 of our 12 participants reported that they had previously leveraged *crowdsourcing* or *in-situ user feedback* approaches to attempt to engage users in surfacing harmful machine behaviors *at scale*. This included using both third-party and internally-built crowdsourcing tools, as well as soliciting in-situ feedback directly from users during everyday interactions with their products and services.

In our semi-structured interviews, some participants reported using third-party crowdsourcing platforms like Amazon's Mechanical Turk or company-internal crowdsourcing platforms to invite a large number of crowd workers to help audit their AI products. For example, P3's team found it useful to conduct focus group workshops to have users audit their NLP applications. However, given the number of languages that their applications cover, their team turned to more scalable methods. For instance, to audit a model intended to detect offensive sentences, P3 deployed a task on Amazon's Mechanical Turk to *"ask crowd workers if they find the sentence offensive or not, and ask them to flag the offensive ones."* This approach helped P3's team uncover a large number of offensive sentences that were previously not predicted as offensive by their language models. Similarly, P5's team paid crowdworkers on a third-party crowdsourcing platform to detect potential biases exhibited in their optical character recognition (OCR) model towards hand-written characters in different source languages.

Inspired by recent cases in which users surfaced biases through their day-to-day interactions with the AI systems (e.g., [21, 80]), four participants developed features that were *seamlessly integrated into their AI systems' user interfaces*, to allow users to provide feedback on potential biases and harmful behaviors in-situ, during their everyday interactions with a system. For example, P7 led an effort to build a plug-in to help users report any harmful outputs encountered while using their image recognition service. Similarly, P10 led an effort to build a web-based application where users could interact with an unreleased conversational agent prototype and report any problematic algorithmic behaviors they encountered during their interactions with the prototype.

	Approaches that do not engage users		User-engaged approaches for testing and auditing their AI systems				
	Fairness toolkits	Internal testing within the product team	User studies	Focus groups	Crowdsourcing through third-party platforms	Crowdsourcing through internal platforms	In-situ user feedback
P1	X	X		X		X	X
P2		X	X				
P3	X	X		X	X		
P4	X		X	X			
P5	X	X	X	X	X	X	
P6		X	X	X			
P7		X	X		X		X
P8	X			X		X	X
P9	X	X	X			X	
P10	X	X	X	X			X
P11	X	X	X		X	X	
P12	X	X		X		X	

Table 2: Participants' existing practices for testing and auditing their AI systems for harmful behaviors and biases. In this table, we report both approaches that do not engage users and user-engaged approaches. Section 4.1.1 describes the *limitations* of approaches that do not engage users. Section 4.1.2 describes the user-engaged approaches in detail. In this table, approaches that do not engage users are color-coded in gray; approaches for user-engagement at small scales are shown in light green; and approaches for user-engagement at larger scales are shown in darker green.

4.2 Challenges and needs in user-engaged algorithm auditing

Participants noted that although their current crowdsourcing and in-situ user feedback approaches could mitigate some of the limitations of existing auditing approaches, they still faced numerous challenges in effectively engaging users beyond existing design considerations for human computation or user feedback systems. For example, detecting harmful algorithmic behaviors requires recruiting and incentivizing the right group of auditors, with relevant cultural backgrounds, lived experiences, and perspectives. In addition, designing auditing tasks can introduce unique complexities, beyond those faced in conventional human computation tasks: the tasks need to guide users towards productive auditing strategies, but without overly influencing them to simply replicate industry practitioners' own biases and blind spots.

4.2.1 ***Identifying and recruiting the right group of auditors.*** All 12 participants emphasized the importance of identifying and recruiting users with relevant identities, cultural backgrounds, and expertise to better test and audit the AI systems.

Identifying relevant subgroups of users: While discussing the design of the *"developer-led"* user-engaged auditing pipeline, many participants (N=6) shared that they found it challenging to determine which demographic subgroups were most critical to engage in auditing specific AI products. For example, P6's team wanted to engage real-world users to audit their conversational AI products, yet they did not know *"who are these users, how much they will be impacted by our product, and how to reach out to them."* When P7's team attempted to recruit users to assist in auditing their image search service, they became overwhelmed as they *"started to think about the intersectionalities [of users], and the demographics just blew up into a billion different categories"* leading P7's team to wonder *"what is the right level of identity intersections to look at?"* P7 noted that this challenge caused their team to *"get lost"* in the process of setting up their own user-engaged auditing pipeline. Similarly, P9, whose team worked on building fairer recommendation systems, shared, *"We were having [a] hard time defining the genre of the content from the artists, and we don't want to just label the artists by ourselves and project our biases [towards] the users even before we started to engage them."* Given this experience, P9 viewed the challenge of identifying relevant subgroups of users as *"a fundamental but intractable first step"* to conducting user-engaged auditing on their product.

Recruiting a diverse and representative set of users: All 12 participants shared that, even when they knew what identities or background expertise they wanted to target, their current approaches were inadequate to actually recruit the targeted groups of users. Six participants reported that, in order to recruit user auditors, they relied heavily on personal networks or existing relationships with previous users. As such, they encountered challenges in recruiting users auditors from demographics, domains, regions and cultures they had not established such relationships with. For example, P1 shared that, as a US-based company with few Asian employees, their team encountered challenges in recruiting *"users from Japan and Korea"* to judge, for example, whether their AI services generate *"potentially offensive labels for a sea area between Japan and Korea, given that these two countries had previous conflicts on the naming and jurisdiction of that area."* Similarly, P12 mainly relied on their team members' existing personal relationships with customers to recruit marginalized community members, but often stumbled when attempting to reach out to a community that none of their team members had previously interacted with.

During the co-design sessions, several participants (N=9) expressed the belief that future tools that support more *"user-led"* approaches to auditing, which appeal to users' *intrinsic* and *social* motivations to participate, could be helpful in reaching users with specific identity characteristics and/or domain knowledge.

For example, P5 drew an analogy to Wikipedia, believing that a more *"user-led"* process could attract users to voluntarily and collaboratively audit AI systems that affect their lives, similar to how *"people collectively edit articles on Wikipedia based on their interests, and will attract people with similar interests to join."* Similarly, P12 saw potential for more *"user-led"* approaches to user-engaged algorithm auditing to organically attract people with similar identities and shared experience: *"For example, people on Reddit with shared identities will come together and discuss problems they are facing."*

4.2.2 ***Incentivizing sustained, high-quality user contributions.*** While previous work has shown both external and intrinsic motivators driving online collective actions [43, 51, 55], in our study, we found that practitioners currently rely primarily upon external motivators, such as financial and social incentives, to motivate user auditors to make sustained, high-quality contributions. Meanwhile, participants shared challenges in employing these motivators.

Challenges in implementing financial incentivization in more "user-led" pipeline: All 12 of our participants shared that they currently provide financial compensation to motivate user auditors. While participants found it relatively straightforward to implement financial incentivization in a more "developer-led" auditing pipeline, during the co-design activities, five participants raised challenges around how to compensate users in a "user-led" auditing pipeline that was more exploratory and discussion-based. For example, P7 commented that the *"open-endedness"* in the "user-led" auditing pipeline also made it difficult to decide how to compensate user auditors: *"How do you pay people in this context? Right? Like what with the task-based things, it's like, there's a clear incentive: you do the tasks, you get paid. Here, where it's exploratory, [do] you pay people for just spending time in this interface? Do you pay them for just chatting? Or just generating hypotheses?"* Similarly, P10 shared a prior experience where their team *"had discussed paying the users who gave good amounts [of] and quality feedback"*, but the conversation died when their team struggled to come up with clear definitions of "good amounts" and "quality" to implement a concrete compensation plan.

Benefits and risks of social motivators: P1's team leveraged social motivators like "peer recognition" and "social interaction" by creating a "star system" to reward high quality auditors, as they found users enjoyed earning stars to demonstrate their "high reputation." P1's team further *"implemented a leader board to keep track of who is bringing in the best feedback, the most feedback, and whose feedback is being endorsed by lots of other auditors,"* and to allow users to write *"recommendation letters"* for one another to audit other tasks. During the co-design study, several participants (N=8) believed that a more "user-led" auditing pipeline could potentially amplify social motivators. Nevertheless, participants raised concerns that *"certain user groups' voices might be further marginalized in [the 'user-led'] pipeline"* (P9). To combat this, participants suggested developers should intervene and facilitate the conversations among the user auditors, to amplify marginalized voices throughout user-engaged auditing processes.

4.2.3 ***Effectively scaffolding users in auditing algorithms.*** Participants noted that it was challenging, in practice, to design user auditing tasks and instructions that could empower users to generate meaningful insights about their AI products and services. As we discuss below, participants shared several challenges they had faced in guiding user auditors without imposing the development team's own biases upon them, and in prompting user auditors to provide more critical feedback on an AI system's overall design.

Guiding users towards productive auditing strategies, without overly biasing them: During the interview portion of our study, participants shared experiences where user auditors had misunderstood the tasks they were given or had failed to provide sufficient detail and context for industry teams to act on their reports. Thus, throughout the co-design portion of our study, several participants (N=8) expressed desires for ways to help user auditors better understand their team's intended goals for an audit, and to scaffold them in auditing a system more effectively. For instance, P5 suggested *"sharing with the users a theoretical structure of biases, an algorithmic harm taxonomy"* to reference both during an onboarding phase and at any point during their auditing activities. However, P5 was uncertain what such a taxonomy would look like in their context (i.e., sentiment analysis). Similarly, P8 and P10 desired better ways to nudge user auditors to *"think out of the box"* (P8) and *"ask hard questions [to a conversational agent] and break the model and surface our 'unknown unknown'"* (P10). P7 noted that, in the context of image search, user auditors may not always test the impacts of small perturbations, so they suggested prompting users to do so: *"Chang[ing] a small word might lead to a very different search result, and we definitely want to guide users to explore these small changes'* (P7).

Despite this desire to guide user auditors in more productive directions, some participants (N=6) expressed concerns that providing too much or the wrong kinds of scaffolding might bias users to think too much like their own teams—potentially limiting the value of a user-engaged auditing approach. For example, during our co-design activity, P10 noted that designing guidelines and specific prompts for user auditors was *"quite tricky since of course we want to offer detailed guidelines and ask specific questions like 'do you think this output is biased towards Asians or women," but our questions might actually bias the users when they are finding biases. [...] We need to make sure that we don't let our confirmation bias affect this [user-engaged audit process]"*.

Acknowledging the challenge of navigating these tradeoffs, five of our participants emphasized the importance of cross-functional collaboration in designing effective guidelines, tasks, and prompts for user auditors. For example, in most of these teams, the design of auditing tasks was left to engineers who had no training in the design of human subjects research methods. By contrast, P10 worked on a team where UX practitioners were involved in the design of auditing tasks. However, the UX and AI teams often worked in silos. P10 felt that *"with a UX background, I only have a surface understanding of large language model[s]. I can't design a good auditing task just by myself if we eventually want to incorporate some auditing feedback from users into the current model."*

Soliciting critical and holistic feedback from user auditors: Ten of our 12 participants expressed desires for better ways to prompt more critical and holistic feedback from user auditors. While iterating on the user-engaged audit report and the "developer-led" auditing pipeline, P9 suggested that in order to prompt critical feedback, it could be helpful to share the team's *rationales* behind particular design decisions, which might otherwise remain opaque

to user auditors. P9 shared a prior experience where their team had initially struggled to gather the sorts of critical feedback they were hoping for. However, after their team shared more details about specific design rationales, *"[user auditors] asked questions like, 'Do I actually want to get recommendations in this way? What's a better way to design this that fits my preference?' These types of questions are the ones we wish our users [would] ask when auditing and include [in] their final report."* Similarly, while describing their team's prior experiences with user-engaged audits, P7 mentioned that users sometimes express a desire to know more about *"why and how"* their AI products and services are designed, in addition to seeing the AI's outputs, so that they could surface potential *procedural* issues that might not otherwise be as visible to them.

Several participants believed that a more "user-led" algorithm auditing process could help to catalyze more critical inquiries from users. For example, while iterating on the "user-led" audit pipeline, P12 said, *"when it's 'developer-led,' we are still testing if there are mistakes or unfairness in the product, right? But what if users believe this AI product shouldn't even exist? This is something you can get by giving people more freedom to discuss."* For this reason, P1 believed it was critical to *"allow [user auditors] to chat with each other,"* in contrast to conventional crowdsourcing approaches, in which crowdworkers perform tasks in independent silos. Similarly, P10 argued that platforms or tools for user-engaged algorithm auditing should include mechanisms *"for users to share and discuss the issue they found during the auditing process"*

4.2.4 ***Deriving actionable insights from user-engaged auditing.*** Seven out of 12 participants shared challenges they currently face in deriving actionable insights from the user audit reports. In particular, participants shared that, unlike in conventional crowdsourcing approaches, understanding the perspectives of "outliers" may often be more important than understanding the majority view. In addition, participants found challenges in communicating qualitative auditing results to key decision-makers, given an organizational culture of valuing numbers over more complex stories.

'It is no longer simply checking the majority vote." Aggregating and interpreting user-engaged auditing reports: Throughout our co-design activities, all 12 participants highlighted the challenges of aggregating and interpreting results from user-engaged auditing processes. As P3 put it, *"in more traditional numerical crowdsourcing activities, you would throw away a person who like always contradicts what everyone else says."* However, P3 noted that the "outliers" in a user-engaged audit are often the ones that developer teams care the most about. These "outliers" may represent users in the margins, who are sensitized to issues that other auditors are not: *"So maybe like everyone said, sample A was not offensive or problematic, except for like, auditor number 39 [...] because number 39 actually found problematic things others didn't"* (P3). Similarly, P2 said that in their view, *"a few [user] audit results stating potential biases and harms might weigh more than one hundred similar good audit results"*, similar to doing UX tests when *"a single negative review might surface key insights for room to improve."*

The challenge of interpreting results in aggregate becomes especially hard when the number of user audit reports gets large. P7 stressed the importance of collecting *"'why' information"* from users, such as open-text responses explaining why a user perceives a particular algorithmic behavior as problematic. However, P7 complained that they currently lack an *"efficient mechanism to combine quantitative and qualitative insights from users' feedback [...] it is no longer simply checking the majority vote."* While discussing the user-engaged audit report, P5 said, *"When I have thousands of training data annotations from crowd workers, I could just check the statistics of the aggregated results [...] but now with thousands of these audit reports, how am I supposed to figure out the most valuable information?"* P5 believed that in order to institutionalize user-engaged auditing in their organization, their team needed to invest in developing new automatic pipelines to augment their current manual process of reviewing user-engaged audit reports.

"It's just our current culture, we still believe more in the numbers": Quantification and its challenges: All 12 of our participants mentioned that, in order to effectively integrate findings from a user-engaged audit, they need to be able to present clear, quantifiable metrics to leadership and other team members. As P11 put it, *"It's just our current culture; we still believe more in the numbers."* However, this often presented challenges in the context of user-engaged algorithm auditing. For example, when P12 could not offer a clean cut number but only *"a complex story"* in response to questions from developers such as *"[what] percentage of users believe their recommendation is bad,"* they were told by their product manager that they were being *"distracting and counterproductive"* to project progress. While iterating on the user-engaged auditing report, P6 shared related concerns about how to measure the progress of a user-engaged audit: *"One consideration for combining these user-engaged reports is, what is the metric to define 'success' here? How many [reports] is enough? How much more do we need before we stop?"* P6 suggested that in order to effectively translate these reports into concrete actions from the product team, defining such metrics and clearly scoping the goal of a user-engaged audit would be critical given that *"[monetary and time] cost is [always a] concern."*

4.3 Organizational obstacles to user-engaged algorithm auditing

Beyond challenges in effectively engaging users, participants also shared broader, organizational obstacles they perceived around potential PR risks, profit motives that work against protecting marginalized groups, and privacy and legal concerns. Taken together, our findings shed light on the ways practitioners currently navigate organizational tensions specifically around user-engagement in algorithm testing and auditing.

4.3.1 ***"What if this destroyed the company's reputation?" Potential PR risk.*** Multiple participants (N=6) raised concerns regarding the feasibility of full institutional buy-in to user-engaged auditing from their organizations. Participants were especially skeptical of the more "user-led" pipeline, as it seemed to hold the greatest potential PR risk. For example, while co-designing the "user-led" pipeline, P7 shared, *"PR issue[s are…] one main reason I see companies don't want [user-led algorithm auditing] as an everyday thing."* P7 backed this fear with a specific experience: *"We had users just try to find the absolute worst thing possible in our models and [they] made it into a story for social media instead of reporting back to us."* P7 also said their team leadership worried that this behavior could *"expose*

the vulnerability of [their] models to their competing companies" and that ultimately *"involving users might create more headlines that damage the company's public image"*. Similarly, P3 said these fears constituted *"a major reason why [their] company is still experimenting [with user-engaged approaches] on some applications instead of making it a company-wide thing."* Even P8, whose company began user-engaged auditing after PR pressure, worried that user-engaged auditing could cause new PR issues. After co-designing the report, P8 said, *"This report would be extremely useful if it only goes to us,"* and wondered how to *"hold the users accountable and make sure they don't destroy the company's reputation after gaining the trust."*

4.3.2 ***"It's all about yelling power": Company motives to amplify certain users' voices over others.*** An important goal of user-engaged algorithm auditing is to translate the problems users find into concrete remediations from the product teams. However, eight of our 12 participants mentioned that, realistically, companies will prioritize addressing issues raised by certain groups of users over others. For instance, P1 said, *"If [...] a large group of researchers from [a major US-based research institute] and a single user from a community college both raised concerns about our knowledge graph, unfortunately our business team would have to prioritize the former."* P1 concluded that when addressing issues raised in user-engaged audits, *"It's all about [the] yelling power of the users."* During the co-design activity, P12 stated that in order to incorporate user-engaged audit in their day-to-day AI work, *"the biggest challenge is not to design the perfect workflow, but to make sure [the company] wants to do [user-engaged auditing] for social good, [not just] for earning more money from more people."* Similarly, P4 shared that in their current user-engaged auditing work, they constantly find themselves battling the business teams since *"the business teams sometimes choose to neglect the users' audit outcome if the reports were not from their 'original [target] audience.'"* P4 believed that, if their team is not implementing remediation in response to reports from the most marginalized users, their organizations *"run the risk of participating in an ethics-washing activity."*

Several participants also shared that they often ran up against a "vicious cycle" (P12), in which a dearth of data from low-resource areas and marginalized communities makes it difficult for practitioners to advocate for more resources to address these areas. P12 said they struggled to get enough resources to test their language technologies with marginalized communities, as their data scientists required large-scale, quantitative evidence before approving studies with new groups of community members. Yet as P12 noted, *"There is, of course, not enough evidence, when these groups were not even considered as users in the first place."*

4.3.3 ***"There are always things we can't get and things we can't share": Privacy and legal concerns.*** As mentioned in Section 4.2.1, participants highlighted that access to user auditors' demographics and other background information is critical in order to assign tasks to appropriate user auditors and to understand which perspectives are represented in reports from user auditors. However, all 12 participants also shared challenges around obtaining certain demographic information due to privacy and legal concerns—mirroring challenges that practitioners face in AI auditing work more broadly [20, 42, 89]. However, beyond standard concerns around the collection and use of sensitive data for AI auditing, participants also shared concerns about the data user auditors might share on an auditing platform, as well as the data they might need to share with user auditors to enable effective audits. For example, when discussing the *"user-led"* auditing pipeline, P7 noted the challenges that could arise if a user auditor were to *"[take] somebody else's photo and share it with other users for the purpose of auditing"* or *"[share] any users' race and gender with other users if they don't want to share."* P10, on the other hand, brought up their internal concerns around *"losing the competition [with other companies] on building large language models,"* by exposing too many model details to user auditors. As P10 shared at the end of the co-design session, *"There are always things we can't get and things we can't share."*

5 LIMITATIONS

Similar to prior HCI work studying responsible AI practices in industry (e.g., [42, 57, 70]), our findings shed light on current practices, challenges, and needs among a set of practitioners who may be at the *forefront* of an emerging industry practice. As discussed in Section 3.1, we recruited participants using a purposive sampling approach [19]. All of our participants were passionate about addressing harmful behaviors in their AI systems, and they had direct experience experimenting with user-engaged approaches to algorithm auditing in their work. Furthermore, most of our participants worked at large technology companies, and all of our participants were located in the US (see Table 1).

6 DISCUSSION

Drawing upon prior literature in areas such as algorithm auditing, crowdsourcing, participatory design, and fairness in AI, we discuss opportunities for future HCI research to help realize the potential and mitigate the risks of user-engaged auditing in industry practice.

6.1 Unique challenges in supporting user-engaged algorithm auditing

Prior research demonstrates great potential for human computation and crowdsourcing approaches to advance AI research and practice, especially in areas such as data generation and annotation [18, 75] and human-level evaluation [5, 88, 93, 94]. In our study, we found that industry practitioners' current approaches to user engagement in algorithm auditing are often built atop existing crowdsourcing pipelines (Section 4.1.2). However, practitioners quickly ran up against limitations of conventional human computation and crowdsourcing approaches (Section 4.2). Below, we highlight five design implications for user-engaged auditing that extend beyond standard considerations for human computation approaches, and discuss corresponding directions for future HCI research.

6.1.1 ***Focusing on "who".*** First, crowdsourcing approaches used to support AI research and practice typically focus more on *what crowd workers do* and less on *who they are* [5, 18, 75, 83, 88]. However, the *"who"* factors that encompass users' intersectional identities and lived experience play critical roles in user-engaged auditing, as discussed in Sections 4.1.1 and 4.2.1. Prior work has also shown that users' personal experiences with and exposures to bias influence

the ways they search for and make sense of harmful behaviors in algorithmic systems [16, 25, 27, 28]. Future research should explore better processes and tools to support practitioners in identifying and recruiting appropriately diverse and representative user auditors, for particular algorithm auditing tasks.

6.1.2 ***Supporting sustained, long-term contributions.*** Second, in contrast to the relatively transient, on-demand nature of most crowd work [12, 50], user-engaged auditing often requires sustained, long-term contributions from user auditors, to continuously improve AI systems (Section 4.2.3). This requirement entails radically different designs on both the process and interface levels. For example, future research may explore the interaction design space of in-situ feedback mechanisms, to solicit user auditors' feedback in the context of their day-to-day interactions with algorithmic systems. A key challenge for this line of exploration is to solicit feedback on algorithmic behavior in formats that are quick and relatively unobtrusive to collect from users, yet at the same time are readily interpretable and actionable for AI practitioners on the receiving end [8, 36]. To better support long-term contributions in user-engaged auditing, it is also critical to design intuitive and efficient interactive interface for users auditors [91]. In addition, future research may explore the design of social platforms to build sustained online communities of users who are motivated to engage in testing and auditing algorithmic systems together [80].

6.1.3 ***Navigating inherent ambiguities in auditing tasks and outcomes.*** Third, traditional crowdsourcing typically starts with well-defined goals and anticipated outcomes set up by the "requesters" [49]. In contrast, there may be benefits to empowering users to collectively take the lead in directing auditing efforts (Section 4.2.3). With too much direction and guidance from industry AI teams, user-engaged audits risk replicating the very biases and blindspots that they were meant to overcome (Section 4.2.3). Prior HCI research has mainly focused on developing tools and processes to better prompt users discovering more "unknown unknowns" [9, 85, 91]. Future research should explore the design of scaffolding mechanisms for user auditors that can navigate the trade-offs between promoting more effective algorithm auditing behaviors versus providing too much direction, limiting the kinds of issues user auditors are able to surface.

6.1.4 ***Reconsidering aggregation and quantification approaches to ensure marginalized voices are heard.*** In addition, as discussed in recent HCI research (e.g., [35]), typical crowdsourcing approaches involve aggregate analyses and evaluations, for example, by relying on "majority vote" from the crowd in order to arrive at the results. Practitioners emphasized that relying exclusively on quantification to derive actionable insights from user-engaged auditing could be harmful and counterproductive (Section 4.2.4). Resonating with findings from prior research [24, 56], we found that practitioners desired practical tools to support them in advocating for marginalized communities through an integration of both qualitative and quantitative forms of evidence (Section 4.3.2). Strategies like "tactical quantification" proposed by Irani et al. [45] could support practitioners in advocating on behalf of user auditors within their organizations, in industry contexts where numbers are culturally valued over more complicated stories. Future HCI research should explore the design of new tools, computational techniques, and visualization approaches that can aid practitioners in persuasively advocating on behalf of marginalized groups of users (cf. [35, 74, 81]).

6.1.5 ***Designing user-engaged auditing mechanisms with teeth.*** Finally, prior work has emphasized the activist nature of auditing, which differs from much prior research on human computation and crowdsourcing [60]. The end goal for user-engaged auditing is to improve products and protect future users from algorithmic harms. Yet achieving this goal requires that companies are actually held accountable for addressing the issues that user auditors uncover [16, 39, 68, 78]. Thus, it is critical for future HCI research to design for user-engaged auditing with accountability and collective empowerment in mind. For example, researchers could explore building platforms that support user auditors in collectively applying pressure and holding companies to account, when serious issues are not addressed [45, 76]. In the next section, we further expand on the discussion around these design implications.

6.2 The complex relationship between industry practitioners and user auditors

Our study explored industry practitioners' perceptions of and relationships with user-engaged approaches to algorithm auditing. A key component of this is the relationship between practitioners and the user auditors who power the auditing process. As these two groups strive to surface and address harmful algorithmic behaviors, a complex relationship and a delicate *mutual (dis)trust* is revealed that can be friendly or antagonistic, depending on the situation.

6.2.1 ***How might users trust AI practitioners?*** For user auditors, auditing is often seen as a form of activism, in which rooting out harmful behaviors in algorithmic systems benefits society [60]. Past research on user-engaged algorithm auditing highlighted users' advocacy for marginalized groups of people, expressing solidarity via their auditing activities [25]. However, industry practitioners in our study rarely brought up similar rationales for engagement with user-engaged auditing (Section 4.2.2). This might be because industry practitioners have greater opportunity to directly effect change on issues in the algorithmic systems they work on, whereas users typically need to rely on their collective power to raise awareness in order to be heard [60, 80]. Though user-engaged audits can empower users through the ability to directly connect with industry practitioners to try and identify issues together, they also firmly place the choice to take action with practitioners, potentially leaving users with less room for leverage via other means.

Given this asymmetric power dynamic between user auditors and AI practitioners, how could HCI researchers support empowering users' collective action when users' needs are not met and their trust fractured? Previous platforms like Turkopticon [45] and WeAreDynamo [76] demonstrated the potential for HCI researchers to consciously build spaces for activism, leveraging their collectiveness to negotiate their desires and needs. Incorporating similar spaces into future user-engaged auditing processes could alleviate users' concerns and empower them to act, ensuring that issues they collectively surface will be addressed in satisfactory ways.

6.2.2 ***How might AI practitioners trust users?*** Despite the desire for holistic and critical auditing processes (Section 4.2.3), industry practitioners in our study expressed trust concerns around opening their systems to be audited by users (Section 4.3.1). As described above, users frequently turn to public awareness raising around problematic algorithmic behaviors they encounter as leverage to pressure companies into addressing those issues. How can practitioners fully trust users to audit their systems without publicizing issues before practitioners have had a chance to address them? Indeed, in our study, practitioners frequently cited apprehensions that user auditors might harm company reputations through negative PR (Section 4.3.1). However, other practitioners viewed themselves as more on the side of users. These practitioners work beyond their main job functions to advocate for users, while simultaneously using the issues brought by user auditors as evidence to convince their teams to act, putting pressure on companies from the inside. How might we support and protect practitioners who genuinely strive to mitigate or avoid problematic algorithmic behaviors in their AI products and services? To this end, similar to supporting user activism, there may be opportunities to build platforms (cf.[45, 76]) to support collective actions amongst these practitioner individual advocates within companies.

One avenue of exploration to address the tensions of required mutual trust might take inspiration from security bug bounties [58]. In these, security experts turn over information about security vulnerabilities to companies in exchange for monetary compensation and the promise of a fix, with the understanding between parties that if the issue is not addressed in a given time frame, then the security expert will publicize the issue. Borrowing from this model could enhance the trust between user auditors and industry practitioners as well, allowing practitioners a protected timeframe to fix issues but enacting a strict deadline for users to take further action if practitioners' promises are not upheld. Indeed, emergent projects like bias bounties [21, 69]have begun to transfer some bug bounty success to the territory of user-engaged algorithm auditing.

6.3 Users' perspectives on user-engaged algorithm auditing

As prior research has begun to explore users' practices and perspectives around user-engaged algorithm auditing [25, 53, 80], our study begins to fill the gap between this emerging literature and industry practitioners' perspectives. However, while supporting practitioners in designing and implementing more effective forms of user-engagement in algorithm auditing, it is critical to continue to explore users' perspectives and values. To this end, future work should bring together industry practitioners' and users' perspectives, potentially through the collaborative design, development, and oversight of user-engaged auditing procedures and platforms.

Importantly, when users are engaged in algorithm testing and auditing, they necessarily observe and likely experience some of those harms themselves. In our study, practitioners described thinking about better ways to target user auditors with relevant identity characteristics. This is in line with prior research, which has found that people with certain exposures and experiences are more able to surface related issues in algorithmic systems [25]. These people may be ideal candidates to serve as user auditors. At the same time, since these are often members of marginalized communities, who are already overburdened and more likely to be the targets of harmful algorithmic behavior [38, 44, 82], they are also more likely to be harmed through participation in algorithm testing and auditing. Furthermore, drawing an analogy to platform content moderation, in which moderators are often exposed to violence and harassment and could be subject to long-term psychological harms [26, 84], auditing for problematic algorithmic behaviors may also result in long-term psychological harms towards user auditors.

Therefore, future research should consider and design to alleviate potential emotional burdens and psychological harms toward user auditors. Furthermore, user-engaged algorithm auditing could harm users if their labors are co-opted in ways that are not aligned with what they might have wanted. We highlight these burdens on users as vital areas for further research. Despite the burdens, user-engaged auditing, when implemented well, can serve to reduce algorithmic harms present and acting in the world now. We urge continual evaluation of user-engaged auditing processes by practitioners to ensure that these burdens and potential harms are mitigated. Future research should also explore the potential of computational or other alternative solutions to reduce the need for continuous involvement of the most vulnerable populations in the auditing process (e.g., [35, 53]), with the caution that computational approaches may also risk introducing new types of harms to users.

7 CONCLUSION

We conducted a series of interviews and iterative co-design activities with industry practitioners to gain insights into the current landscape and future opportunities for user-engaged algorithm auditing in industry practice. We surfaced major motivations for engaging users in AI testing and auditing and described practitioners' existing approaches for user-engaged auditing. We found that practitioners face challenges around appropriately recruiting and incentivizing user auditors, scaffolding user audits, and deriving actionable insights from audit reports. Furthermore, practitioners shared broader organizational obstacles to user-engaged auditing, highlighting key tensions that arise in practice when involving users in algorithm auditing efforts. Based on these findings, we discussed the complex relationships between practitioners and user auditors, offering potential remediation for developing mutual trust. We then describe various opportunities to support user-engaged auditing beyond existing design considerations for human computation or user feedback systems. Overall, we hope that this work inspires future efforts to realize the potential and mitigate the risks of user-engaged auditing in industry practice.

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 award from Cisco Research, and an award from the Jacobs Foundation. We would like to thank Alex Cabrera, Charvi Rastogi, Tzu-sheng Kuo, Kimi Wenzel, Seyun Kim, Katelyn Morrison, Adam Perer, Jason Hong for their feedback on the draft. Special thanks to our anonymous reviewers and to all participating industry practitioners for making this work possible.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 252–260.

[2] IBM Resaerch Trusted AI. 2021. AIF360 API. (2021). https://aif360.mybluemix.net/

[3] IBM Resaerch Trusted AI. 2021. AIX360 API. (2021). https://aix360.mybluemix.net/

[4] Open AI. 2022. ChatGPT Feedback Contest: Official Rules. https://cdn.openai.com/chatgpt/ChatGPT_Feedback_Contest_Rules.pdf

[5] Dimitra Anastasiou and Rajat Gupta. 2011. Comparison of crowdsourcing translation with Machine Translation. *Journal of Information Science* 37, 6 (2011), 637–659.

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[7] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing race and gender discrimination in online housing markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 24–35.

[8] Joshua Attenberg, Panos Ipeirotis, and Foster Provost. 2015. Beat the machine: Challenging humans to find a predictive model's "unknown unknowns". *Journal of Data and Information Quality (JDIQ)* 6, 1 (2015), 1–17.

[9] Josh M Attenberg, Pagagiotis G Ipeirotis, and Foster Provost. 2011. Beat the machine: Challenging workers to find the unknown unknowns. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[10] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[11] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[12] Jeffrey P Bigham, Michael S Bernstein, and Eytan Adar. 2015. Human-computer interaction and collective intelligence. *Handbook of collective intelligence* 57 (2015).

[13] Sarah Bird. 2020. Fairlearn API. https://fairlearn.github.io/v0.5.0/api_reference/fairlearn.datasets.html

[14] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[15] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[16] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[17] Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. 2021. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–22.

[18] Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*. 1–12.

[19] Steve Campbell, Melanie Greenwood, Sarah Prior, Toniele Shearer, Kerrie Walkem, Sarah Young, Danielle Bywaters, and Kim Walker. 2020. Purposive sampling: complex or simple? Research case examples. *Journal of research in Nursing* 25, 8 (2020), 652–661.

[20] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*. 339–348.

[21] Rumman Chowdhury and Jutta Williams. 2021. Introducing Twitter's first algorithmic bias bounty challenge. *URl: https://blog. twitter. com/engineering/en_us/topics/insights/2021/algorithmic-bias-bountychallenge* (2021).

[22] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and addressing algorithmic bias in practice. *Interactions* 25, 6 (2018), 58–63.

[23] Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

[24] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 473–484. https://doi.org/10.1145/3531146.3533113

[25] 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. *CHI Conference on Human Factors in Computing Systems* (2022).

[26] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[27] 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

[28] 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

[29] 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.

[30] 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. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.

[31] 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

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

[33] Shelley Evenson. 2006. Directed storytelling: Interpreting experience for design. *Design Studies: Theory and research in graphic design* (2006), 231–240.

[34] 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

[35] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[36] Alex Groce, Todd Kulesza, Chaoqiang Zhang, Shalini Shamasunder, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Shubhomoy Das, Amber Shinsel, Forrest Bice, et al. 2013. You are the only possible oracle: Effective test selection for end users of interactive machine learning systems. *IEEE Transactions on Software Engineering* 40, 3 (2013), 307–323.

[37] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*. 305–318.

[38] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[39] Jamie Harris. 2022. Facebook forced to ban its AI after it 'revealed how to make napalm and made racist comments. https://www.the-sun.com/tech/6729391/meta-withdraws-ai-galactica-controversy/

[40] 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.

[41] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019. Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. (2019), 157–171.

[42] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[43] Mokter Hossain. 2012. Users' motivation to participate in online crowdsourcing platforms. In *2012 International Conference on Innovation Management and Technology Research*. IEEE, 310–315.

[44] Yen-Chia Hsu, Himanshu Verma, Andrea Mauri, Illah Nourbakhsh, Alessandro Bozzon, et al. 2022. Empowering local communities using artificial intelligence. *Patterns* 3, 3 (2022), 100449.

[45] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.

[46] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk. (2011).

[47] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[48] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337* (2021).

[49] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1301–1318.

[50] Aniket Kittur, Jeffrey V. Nickerson, Michael S. Bernstein, Elizabeth Gerber, Aaron D. Shaw, John Zimmerman, Matthew Lease, and John Joseph Horton. 2013. The future of crowd work. *Proceedings of the 2013 conference on Computer supported cooperative work* (2013).

[51] Robert E Kraut and Paul Resnick. 2011. Encouraging contribution to online communities. *Building successful online communities: Evidence-based social design* (2011), 21–76.

[52] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165 (2016), 633.

[53] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 512 (Nov 2022), 34 pages. https://doi.org/10.1145/3555625

[54] Michelle Seng Ah Lee and Jatinder Singh. 2020. The Landscape and Gaps in Open Source Fairness Toolkits. *Available at SSRN* (2020).

[55] Kimberly Ling, Gerard Beenen, Pamela Ludford, Xiaoqing Wang, Klarissa Chang, Xin Li, Dan Cosley, Dan Frankowski, Loren Terveen, Al Mamunur Rashid, et al. 2005. Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication* 10, 4 (2005), 00–00.

[56] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *arXiv preprint arXiv:2112.05675* (2021).

[57] 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. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[58] Suresh S Malladi and Hemang C Subramanian. 2019. Bug bounty programs for cybersecurity: Practices, issues, and recommendations. *IEEE Software* 37, 1 (2019), 31–39.

[59] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[60] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[61] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[62] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

[63] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6. 126–135.

[64] Rodrigo Ochigame and Katherine Ye. 2021. Search Atlas: Visualizing Divergent Search Results Across Geopolitical Borders. In *Designing Interactive Systems Conference 2021*. 1970–1983.

[65] Giada Pistilli. 2022. HuggingFace announcedthe new feature to flag any Model, Dataset, or Space on the Hub. https://twitter.com/GiadaPistilli/status/1571865167092396033?s=20&t=LRhhEu63s6ftPmtZdfz8Cw

[66] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications* 32, 10 (2020), 6363–6381.

[67] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

[68] 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. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[69] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 557–571.

[70] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[71] Microsoft Research. 2022. AI Fairness Checklist. https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/

[72] People + AI Research. 2021. People AI Guidebook. (2021). https://pair.withgoogle.com/guidebook/

[73] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. 2021. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[74] Samantha Robertson and Niloufar Salehi. 2020. What If I Don't Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design. In *Workshop on Participatory Approaches to Machine Learning at ICML 2020*.

[75] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[76] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, and Kristy Milland. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1621–1630.

[77] 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).

[78] Oscar Schwartz. 2019. Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation The bot learned language from people on Twitter—but it also learned values. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

[79] 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

[80] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[81] Hong Shen, Leijie Wang, Wesley H Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 440–451.

[82] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423* (2020).

[83] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.

[84] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[85] 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

[86] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10–29.

[87] Kush R Varshney. 2019. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 26–29.

[88] Jennifer Wortman Vaughan. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.* 18, 1 (2017), 7026–7071.

[89] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017), 2053951717743530.

[90] Tris Warkentin and Josh Woodward. 2022. AI Test Kitchen. https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/

[91] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 747–763.

[92] Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology* 21, 2 (2019), 89–103.

[93] Omar Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 1220–1229.

[94] Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. 2019. Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems* 32 (2019).

[95] 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 APPENDIX

Figure 3: A potential user-engaged audit report that was iteratively co-designed with participants. During the co-design activity, participants could zoom in, annotate, and modify the details. We used this report template as a probe, not as final products, to investigate more deeply on practitioners' challenges and desires.