# Qeios

Research Article

WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI

Wesley Hanwen Deng[1], Claire Wang[2], Howard Ziyu Han[1], Jason I. Hong[1], Kenneth Holstein[1], Motahhare Eslami[1]

1. Carnegie Mellon University, United States; 2. University of California, Berkeley, United States

There has been growing interest from both practitioners and researchers in engaging end users in AI auditing, to draw upon users' unique knowledge and lived experiences. However, we know little about how to effectively scaffold end users in auditing in ways that can generate actionable insights for AI practitioners. Through formative studies with both users and AI practitioners, we first identified a set of design goals to support user-engaged AI auditing. We then developed *WeAudit*, a workflow and system that supports end users in auditing AI both individually and collectively. We evaluated *WeAudit* through a three-week user study with user auditors and interviews with industry Generative AI practitioners. Our findings offer insights into how *WeAudit* supports users in noticing and reflecting upon potential AI harms and in articulating their findings in ways that industry practitioners can act upon. Based on our observations and feedback from both users and practitioners, we identify several opportunities to better support user engagement in AI auditing processes. We discuss implications for future research to support effective and responsible user engagement in AI auditing.

Jason I. Hong, Kenneth Holstein, and Motahhare Eslami equally contributed to this work.
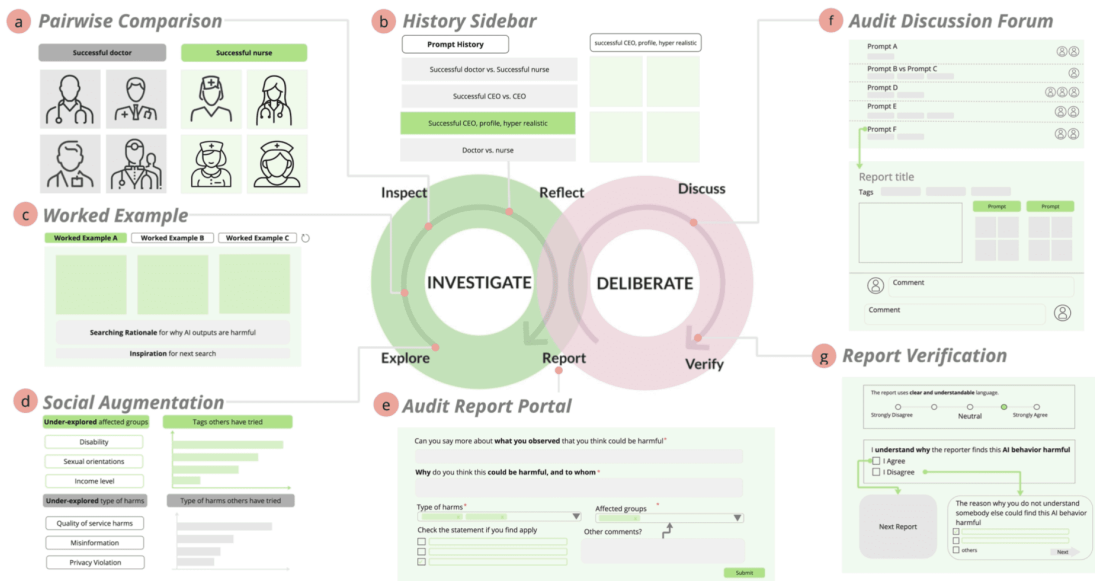
Figure 1. *WeAudit* supports end users in iteratively investigating, deliberating on, and reporting perceived harms and biases in generative AI (GenAI) systems through support for comparing and making sense of different AI generated outputs (a, b); support for reflection around their own and others' auditing findings (b, c, d); and mechanisms for structured reporting, discussion, and verification of audit findings (e, f, g). Through user studies with both end users and industry GenAI practitioners, our work highlights design considerations, recommendations, and opportunities to support effective and responsible user engagement in AI auditing.

1. Introduction

End users often surface harmful behaviors in Artificial Intelligence (AI) systems that are otherwise overlooked by AI practitioners[1][2][3][4][5][6][7][8][9][10][11]. Past research shows that AI users can detect and raise awareness of biased and harmful AI behaviors by leveraging their lived experiences, building on one anothers' findings, and generating and testing hypotheses about AI harms[1][2]. There has been rapidly growing interest among both researchers and practitioners in engaging end users in auditing AI products and services: one line of recent research has explored systems and processes to enable user engagement in auditing[12][13][14][15][16][3][17], while another has focused on understanding industry AI practitioners' needs and challenges around effective user engagement[7][18][19]. However, a critical gap remains in connecting the needs, challenges, and perspectives of these

two groups of stakeholders. **How might we develop tools and processes to effectively scaffold end-user engagement in AI audits, while ensuring their findings are useful and actionable for AI practitioners?**

To bridge this gap, in this work, we introduce *WeAudit*, a workflow and platform, designed with and for both end users and industry AI practitioners, to support meaningful user engagement in AI audit. In line with prior research, throughout this paper we use the term "user-engaged AI auditing" (or simply "user auditing") to refer to processes where users are actively engaged in AI auditing processes that may also include auditing efforts from AI practitioners or other technical experts[7][20][3][21]. We use the term "user auditors" to refer to AI users who are engaged in auditing processes. We focus in particular on Generative AI (GenAI) systems, which have garnered significant attention for user auditing due to their widespread availability and their ability to flexibly generate content over a vast input-output space[22][23]. These characteristics of GenAI introduce an equally vast space of risks and use cases, prompting calls across government, academia, and industry to engage diverse expertise in AI auditing[24][25][26][27][28][29][30].

To understand both end users' and AI practitioners' needs for support in user auditing, we first conducted a formative study with 11 end users, and seven industry AI practitioners with prior experience engaging end users in AI auditing. These formative studies led to the development of a set of design goals, and a corresponding *WeAudit* workflow. We then instantiated the *WeAudit* workflow through a system that supports and scaffolds users in investigating, deliberating, and reporting perceived harms and biases in text-to-image (T2I) AI systems. To understand the strengths and limitations of *WeAudit* in supporting user auditing, we first conducted a three-week user study with 45 users to audit Stable Diffusion, an open source T2I system, and reflect on their experience using the tool. We then conducted semi-structured interviews with 10 industry GenAI practitioners to evaluate the design of *WeAudit*, grounded in data and user audit reports from our user study.

Through these studies, we provide design insights into how *WeAudit* (1) augments user auditors' abilities to detect and reflect on AI biases and harms that might otherwise be overlooked, (2) enhances the depth and breadth of user audits, (3) supports the creation of user audit reports and discussions that AI practitioners find actionable for mitigating AI harms, and (4) helps users increase their awareness and understanding of AI harms. Practitioners found the audit outcomes insightful and could envision incorporating *WeAudit* into their existing GenAI evaluation and development workflows. Our study also revealed several opportunities to improve *WeAudit* and future user-engaged

AI audit processes. We discuss implications of our findings for future research and practice on user engagement in AI auditing. Overall, this work makes the following contributions (Figure 2):

- We identify a set of **design goals** for systems and processes to support user-engaged AI auditing, informed by formative studies with both end users and industry GenAI practitioners.
- We present *WeAudit*, a general workflow and interactive system for user-engaged AI auditing, based on our design goals.
- We share empirical **insights** into how *WeAudit* supports users in auditing GenAI and how industry GenAI practitioners envision adapting *WeAudit* to support their current workflows.
- We discuss the **broader implications** of our findings for HCI and AI researchers and practitioners, organizations, and policymakers in sustaining user engagement in AI auditing and red teaming in a responsible and ethical manner.
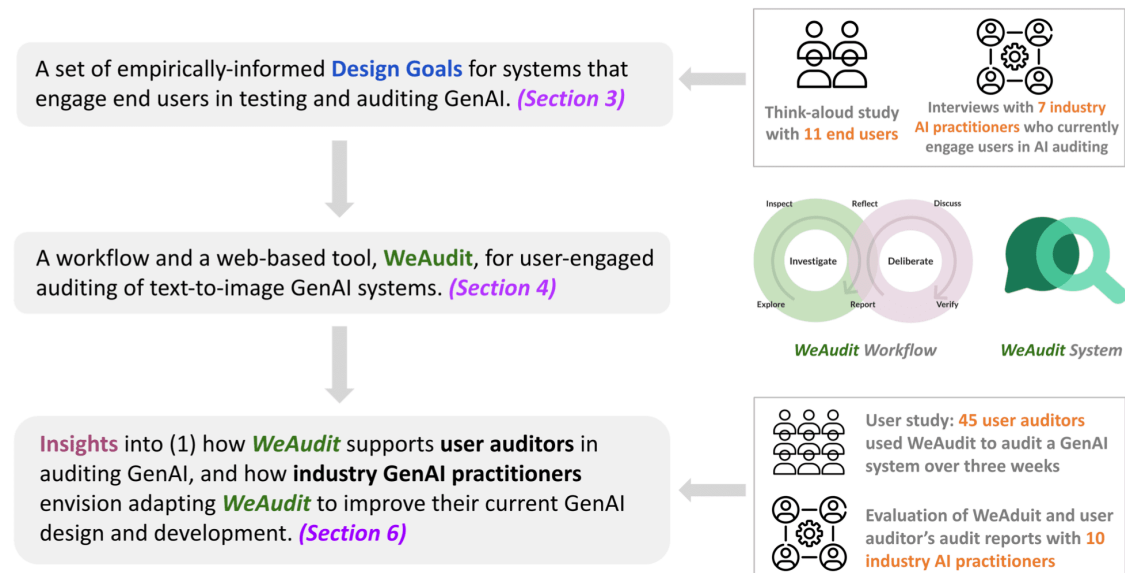

A set of empirically-informed **Design Goals** for systems that engage end users in testing and auditing GenAI. *(Section 3)*

Think-aloud study with **11 end users**

Interviews with **7 industry AI practitioners** who currently engage users in AI auditing

A workflow and a web-based tool, **WeAudit**, for user-engaged auditing of text-to-image GenAI systems. *(Section 4)*

Inspect   Reflect   Discuss
Investigate   Deliberate
Explore   Report   Verify

WeAudit *Workflow*   **WeAudit** *System*

Insights into (1) how *WeAudit* supports **user auditors** in auditing GenAI, and how **industry GenAI practitioners** envision adapting *WeAudit* to improve their current GenAI design and development. *(Section 6)*

User study: **45 user auditors** used WeAudit to audit a GenAI system over three weeks

Evaluation of WeAduit and user auditor's audit reports with **10 industry AI practitioners**

Figure 2. An overview of the methods and contribution of our work.

2. Related Work

2.1. Engaging End Users in Auditing AI Systems for Harmful Behaviors

In recent years, AI audits have gained prominence as a method for uncovering biased, discriminatory, or otherwise harmful behaviors in algorithmic systems[31][32][33][34][35][36][37][38][39][20]. At a high

level, AI auditing refers to a process of repeatedly testing an algorithm with inputs and observing the corresponding outputs, in order to understand its behavior and potential external impacts[39][20].

While most AI audits are still conducted by experts such as researchers[35][40][41][42][43][44] and AI practitioners[33][38][39][45], end users can often identify and raise awareness of harmful AI behaviors in systems that impact their lives, complementing expert-led AI audit efforts[1][2]. For example, end-users often discover harmful biases in Text to Image (T2I) generative AI systems that expert auditors fail to detect[4][46][5][47]. When Mack et al. engaged people with disabilities to review and reflect on images generated by T2I systems they uncovered several societal stereotypes, such as "perpetuating broader narratives in society around disabled people as primarily using wheelchairs, being sad and lonely, incapable, and inactive"[4].

Recognizing the power of users in AI auditing, researchers in CSCW, HCI, and AI have begun exploring tools and processes to support more user-engaged approaches to AI auditing, where users actively participate in the auditing process alongside efforts from AI practitioners or other technical experts[1][2][3][12][13][14][15][9]. For example, DeVrio et al. conducted behavioral studies to investigate how users, both individually and collectively, are able to surface harmful algorithmic behaviors that formal or expert-led audits may miss[2]. Lam et al. developed an "End-user Audits" framework, instantiated in a system called IndieLabel, to engage non-technical users in auditing sentiment analysis classification AI models[3].

In parallel, several major technology companies have begun experimenting with user-engaged auditing approaches to identify harmful and biased outputs in their AI products and services[48][49][50][45][14]. In 2021, Twitter launched its first "algorithmic bias bounty" challenge, inviting users to identify harmful biases in its image-cropping algorithm[45]. Following this, OpenAI introduced a "Feedback Contest" encouraging users to provide feedback on problematic model outputs during their interactions with ChatGPT. Many corporations with generative AI products have also turned to "red-teaming" events, where users rigorously test models to identify vulnerabilities and harmful biases (such as the AI village at DEF CON, a hacking conference)[51][52]. In response to calls from government agencies like the Executive Order on Safe, Secure, and Trustworthy Development and Use of AI[25], many civil society organizations has called for technology companies to engage external domain experts, including end users, in auditing their generative AI products[53][54].

A small but growing body of research has begun exploring the needs and challenges faced by AI practitioners in conducting user-engaged AI audits. For example, through interviews with 35 AI audit practitioners, Ojewale et al. identified the need for tools to support AI practitioners in engaging those directly impacted by AI deployments, offering insights into how to better design participatory AI tools and processes with accountability[18]. In a study with industry AI practitioners who are engaging users in auditing their AI systems, Deng et al. found that practitioners often repurpose crowdsourcing platforms for user audits but struggle to scaffold users toward productive auditing strategies and to derive actionable insights from users' audit outcomes[7].

Despite these two growing lines of research—supporting user audits through tools and processes, and understanding the needs and challenges of industry AI practitioners in user audits—gaps remain in understanding the design goals and in creating tools that **support both AI users and AI practitioners** in the user-engaged AI audit process. Our work extends this prior research by designing, prototyping, and evaluating a user-engaged AI audit workflow and tool that **effectively scaffolds end-user engagement in AI audits, while ensuring their findings are useful and actionable for AI practitioners**. Through evaluating *WeAudit* with both AI users and practitioners, our work provides insights for future practitioners and researchers to better engage users in the AI auditing process.

2.2. Scaffolding Mechanisms for Crowdsourcing

CSCW and HCI research has a long tradition of designing scaffolding mechanisms—including various types of instructions, interfaces, and incentives—to support the crowdsourcing and sensemaking processes for diverse tasks[55][56][57][58]. This line of research explores methods to help crowd workers perform creative and open-ended tasks beyond simple voting or preference selection[59][60][61][56]. For instance, Dow et al. and Chung et al. demonstrated how structured feedback can guide workers in complex tasks, improving the quality of creative contributions[60][61]. Chilton et al. developed a workflow to assist crowd workers in subjective categorization tasks through iterative feedback, highlighting crowds' potential in complex sensemaking[59].

Further studies examine how presenting examples can scaffold crowd work, especially in unfamiliar tasks requiring iterative work and hypothesis generation[62][63][64][65][66][67]. Social augmentation, such as showing others' work, has been shown to improve task completion, as demonstrated by Morris's SearchTogether tool, which enables collaborative search tasks[62]. Other research indicates that timely exposure to diverse sets of examples can stimulate new directions of thought for crowd

workers[64], potentially enhancing the quantity and quality of ideas generated[63]. However, this research consistently highlights that the effectiveness of examples varies based on task content and presentation style[64]. Fianlly, research on reliability in crowdsourced work developed methods for verifying others' contributions, such as Bernstein et al.'s "Find, Fix, Verify" workflow, which enables structured peer review[68].

Drawing from prior crowdsourcing and other CSCW research, our work extends this line of research by designing, prototyping, and evaluating scaffolding mechanisms **specifically for AI users engaged in auditing potentially harmful AI outputs, both individually and collectively.**

3. Formative Study

In this section, we first describe the procedure of our formative study. We then present six Design Goals for processes and tools to support user-engaged AI auditing, informed by our formative study with AI users and practitioners.

3.1. Method

3.1.1. Think-aloud study with end users

To understand the challenges end users encounter when participating in GenAI audits, and to identify opportunities to better support them, we first conducted think-alouds with 11 end users. For these formative studies, we aimed to recruit participants spanning diverse demographics and levels of technical literacy. To this end, we recruited participants through Craigslist and Nextdoor. 10 of our 11 participants had not previously used any T2I systesm; participants ranged from 18 years old to 80 years old, with 47.45 average. 2 participants has high school degrees, 2 Associate degrees, 5 bachelor degrees, 1 master degree, and 1 Ph.D. We include detailed participants descriptions in Table 3 under the Appendix A.1.

In each study, we first shared the goal of the study and then onboarded participants by providing examples of harmful text-to-image results along with reasons why these results could be considered harmful towards certain social groups. These examples and rationales were drawn from prior expert-led audits on T2I systems, covering various common types of T2I harms, such as representational harms and privacy violations.[69][70][71]. Participants then engaged in an initial 20-30 minute model testing phase where they were asked to test a text-to-image model (Stable Diffusion) for harmful

biases, using an off-the-shelf interface provided by Dream Studio[1]. This simple interface allowed users to enter a prompt into a text box and generate 8 images. We encouraged participants to think aloud while brainstorming prompts and inspecting model outputs, and to explain whether and why they found the results potentially harmful.

After this initial phase, for the next 20 minutes participants continued their testing, but now with access to a set of three low-fidelity prototypes intended to support their process. The first prototype was a pairwise comparison feature, simulated by placing two browser windows side-by-side. We hypothesized that viewing prompt outputs side-by-side would help participants, especially those unfamiliar with T2I systems, understand how prompt changes affect the output. The second prototype was a sidebar displaying a sample user audit report in text format. This report featured an election-themed prompt and image pair, with a brief sentence explaining why the AI-generated image could be harmful. We hypothesized that this example could inspire participants to search for similar cases and better articulate the harms they identified. The third prototype was a T2I harm taxonomy, mapping Shelby et al.'s sociotechnical AI harm taxonomy[72] to concrete examples of harmful T2I behaviors, based on examples from[70]. We hypothesized that this taxonomy would help users brainstorm a broader range of potential harms to various communities and use terminology to describe these harms effectively. We include screenshots of each prototype in Figure 3.
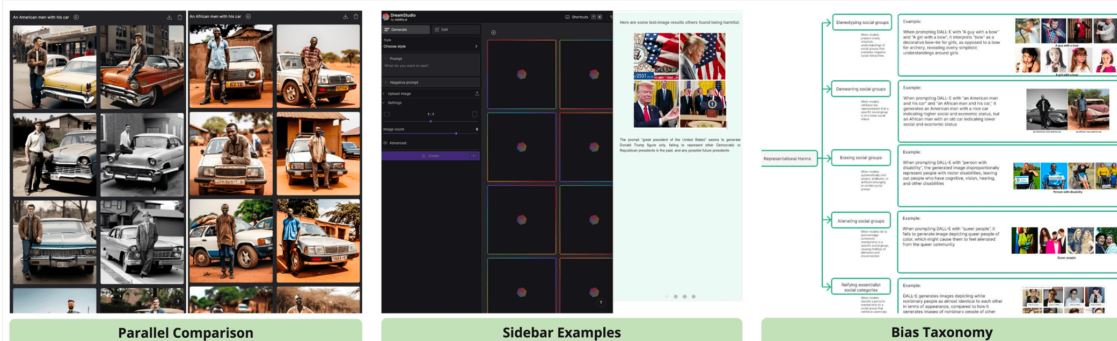

Figure 3. Screenshot of three low-fidelity prototypes we used in our formative study.

3.1.2. Semi-structured Interview with Industry GenAI Practitioners

To understand industry GenAI developers' needs for support in effectively engaging end-users in AI auditing, we conducted semi-structured interviews with 7 industry AI developers. To recruit

participants, we adopted a purposive sampling approach[73], with the aim of recruiting industry practitioners who had already attempted to engage end users in evaluating and auditing their GenAI products[2]. We recruited participants through social media (e.g., LinkedIn and X), and via direct contacts at large technology companies. Overall, industry practitioners come from diverse backgrounds, working in technical, user-facing, and managerial roles on various types of AI products powered by Large Language Models (LLMs) or other types of generative AI. Notably, all of them are currently using crowdsourcing platforms to engage individuals outside their AI teams in auditing or red-teaming their AI products and services. Table 5 in the Appendix A.1 provides an overview of participants' backgrounds[3].

In each interview session, we began by asking participants to describe their current motivations and practices for engaging end users. We then delved deeper into the challenges and pitfalls practitioners had encountered when attempting to engage users in the auditing process. The interview concluded with a discussion of features that practitioners would ideally want in a user-engaged AI auditing workflow.

3.2. Design Goals

Through our formative study with end users (U01 - U11, referred to as "users" throughout this section) and industry AI practitioners (P01 - P07, referred to as "practitioners"), we synthesized the following six design goals for processes and tools to support user-engaged AI auditing.

DG1: Facilitate effective generation and validation of hypotheses about harmful AI behaviors.

We observed that users often iteratively explored small variations on their prompts. For example, after reviewing the results of "An African man and his car," U01 hypothesized that there might be biases in how objects are presented in connection to people, so they continued testing variations of this prompt, such as "An African man and his car in the U.S." and eventually "An African man and his car in Africa." All users reported finding side-by-side comparisons helpful in supporting such exploration. U05 and U10 were interested in further support for making sense of AI-generated images across *many* different prompts, to validate hypotheses or generate new ones.

Similarly, all practitioners expressed desires for interfaces that could support users in more effectively generating and testing hypotheses about potential harmful behaviors exhibited in generative AI. For

example, similar to our observations in the think-aloud study with users, P04 observed in past focus groups with users that when they encountered specific harmful outputs, they often formed hypotheses about broader biases that might be present, and then wanted to try out a similar input to test these hypotheses.

DG2: Support and encourage user auditors to incorporate their lived experiences and identities.

We observed that users often found harmful AI outputs when they began reflecting on their own lived experiences and identities. For example, U02 found more to comment on when testing prompts related to places they had lived, and U06 noted it was easier to spot biases in images related to familiar occupations and activities. Similarly, U07 generated insightful comments by incorporating their hobbies into prompts. However, most pilot study users (with the exception of U06) did not naturally draw upon their personal experiences and identities before we explicitly instructed them to do so, suggesting opportunities for interventions to encourage user auditors to do so.

In line with these observations, all practitioners emphasized the importance of incorporating end users' personal experiences and identities into the auditing process. P01, P02, P04, P06, and P07 all stated that leveraging users' diverse lived experiences to cover the blind spots of their companies' AI teams was a "*key motivation to engage end users*" (P04) but one that their teams currently struggled with. For instance, P06 shared that their team had designated a meeting to strategize ways to better incorporate users' identities into the auditing process, such as providing instructions for users to specifically consider real-life scenarios where they would use the company's AI products.

DG3: Provide appropriate scaffolding to support user auditors in creatively and efficiently auditing AI systems, without overly priming them towards narrow auditing directions.

While the examples we provided to users helped them get started with the auditing process, we observed that some participants over-relied on the provided examples. For example, during the think-aloud session, U03 and U08 only checked the first few examples we provided in the T2I harm taxonomy and example audit report, and the prompts they explored were largely based on those examples. Towards the end of the think-aloud, U06 shared that they felt the T2I harm taxonomy we shared with them could potentially "*limit the kinds of responses... [as it] might shut off some people [who]*

are going to come up with on their own [prompts]… if they feel like they're being shoved into a certain category in their response, it could stifle their creativity."

Practitioners' comments resonated with these observations. For example, P05 shared their observations that providing a small set of static examples often led to users conducting audits that were similar to those examples. P05 further commented, "*It is important to provide users with sufficient scaffolding, especially if they haven't done this type of activity before… but we also want to make sure that they are thinking outside the box and find creative ways to break the model.*" Similarly, P01 suggested that the onboarding and scaffolding materials in the user audit process should not "*overly prime users towards the assumptions held by the development team… good [user auditing] should be a creative endeavor and independent from what developers already know.*"

DG4: Structure the audit report process to allow users to clearly communicate observed harms in a way that helps AI developers extract actionable insights.

During the think-aloud study, we observed all users struggled to clearly articulate which part of the AI-generated content was harmful and why, even when they sensed that something was off. For example, U10 compared the prompts "Rich People" vs. "Poor People". Upon observing that the generated images of "poor people" all shared the same style of dress, U10 was unsure how to describe why they found this problematic. Multiple users (U04, U05, U10, U11) explicitly asked for us to tell them more about what information might be useful for us, suggesting opportunities to elicit their perceptions in more structured ways.

All practitioners mentioned that scaffolding users in crafting reports is essential for them to translate users' feedback into concrete improvements. In particular, practitioners suggested that they would like to know *what users actually observed* in the AI output that they thought could be harmful, *why* they believed this harmful, and *to whom*. To aid in interpreting users' reports, practitioners were also interested in knowing additional information about the user auditor who wrote a given report. For example, all practitioners suggested that audit report could solicit if user auditors come from a specific marginalized background could help them prioritize which audit outcomes to incorporate.

DG5: Enable users to discuss and deliberate on AI harms and mitigation strategies.

Users expressed desire to discuss their findings with others to better explore potential harms or confirm their perceptions of harms toward specific communities that they were less personally

familiar with. For example, when U11 observed the AI-generated output from the prompt "fashion," they wished to discuss it with others, saying, "*some people might say the result is indeed about fashion, some might say the results are all white women, and some might say they are unhealthy-looking women... it might be interesting to hear different comments like that.*" Furthermore, towards the end of the think-aloud session, some users asked if they could review audit reports from others and learn from their auditing strategies.

Many practitioners (5/7) also expressed the importance of facilitating discussions and deliberations among users. For example, P04 mentioned that they conduct focus groups with groups of 4-6 users when auditing their AI products because "*users can often inspire each other to find something new through going back and forth... some users will recall issues they had when someone else mentioned a situation they experienced.*" P01 shared that their team often found it challenging to aggregate and appropriately resolve disagreements in users' Likert scale ratings regarding perceived biases and harms from chatbots toward certain communities. As a result, P01's team was exploring ways to connect end users who submitted conflicting audit reports, allowing them to directly discuss their disagreements so the team could better understand the rationales behind their reports.

DG6: Support careful verification of audit reports without disempowering minority voices.

Finally, all practitioners emphasized the importance of verifying users' audit reports to separate signal from noise when interpreting and prioritizing among a large amount of user feedback. In line with prior work[7][74], practitioners shared that, unlike aggregating the results of traditional AI labeling work, verifying audit outcomes is challenging because: (1) developer teams do not have ground truth, so the verification process must rely on other auditors; and (2) outliers in the data might be important signals, especially if they come from marginalized voices. For example, P05, who had been re-purposing crowdsourcing tools to support user-engaged auditing of generative AI applications, noted, "*quality control in this kind of [AI auditing] task is even harder because [...] we don't have a ground truth either, and everything [crowd workers] surface is new to us.*" P02 shared that a significant challenge their team currently faces is determining whether the representational harms identified by users resonate with others. To this end, all practitioners agreed that verifying audit outcomes is challenging yet critical for the design of a successful user-engaged AI auditing process.

4. WeAudit

Based on our Design Goals, we first synthesized the *WeAudit* workflow, which includes two iterative loops and six sub-activities that is generalizable for scaffolding user engagement in AI audit. We then instantiated the *WeAudit* workflow through a web-based prototype system, with each core feature mapped to the Design Goals and the high-level *WeAudit* workflow. We describe the *WeAudit* workflow and system in the following sections.

4.1. The WeAudit Workflow

As shown in Figure 4, the *WeAudit* workflow is organized into two intersecting, iterative loops: *Investigate* and *Deliberate*. The overall loop structure of this workflow is informed by existing models of information foraging and sensemaking[75], and these loops are further broken down into six sub-activities: *Explore*, *Inspect*, *Reflect*, *Report*, *Discuss*, and *Verify*. In the *Investigate* loop, user auditors may begin investigating AI harms by *exploring* prompts that could lead to harmful outputs (**DG1**, **DG3**). Auditors *inspect* the resulting AI-generated outputs (**DG1**), and *reflect* on potential biases or harmful impacts, drawing upon their unique knowledge and experiences (**DG2**). Auditors can then *report* their findings, documenting potential harms they identified (**DG4**). In the *Deliberate* loop, user auditors can *discuss* their findings with other auditors or the broader public (**DG5**). This collaboration allows for a more holistic understanding of AI-related harms, leveraging user auditors' diverse perspectives. Additionally, reported findings can be *verified* by other auditors or the public (**DG6**), to provide a sense of whether a given observation is perceived as potential harmful by multiple people. In Table 1 above, we further illustrate how each component of the *WeAudit* workflow reflects the Design Goals.
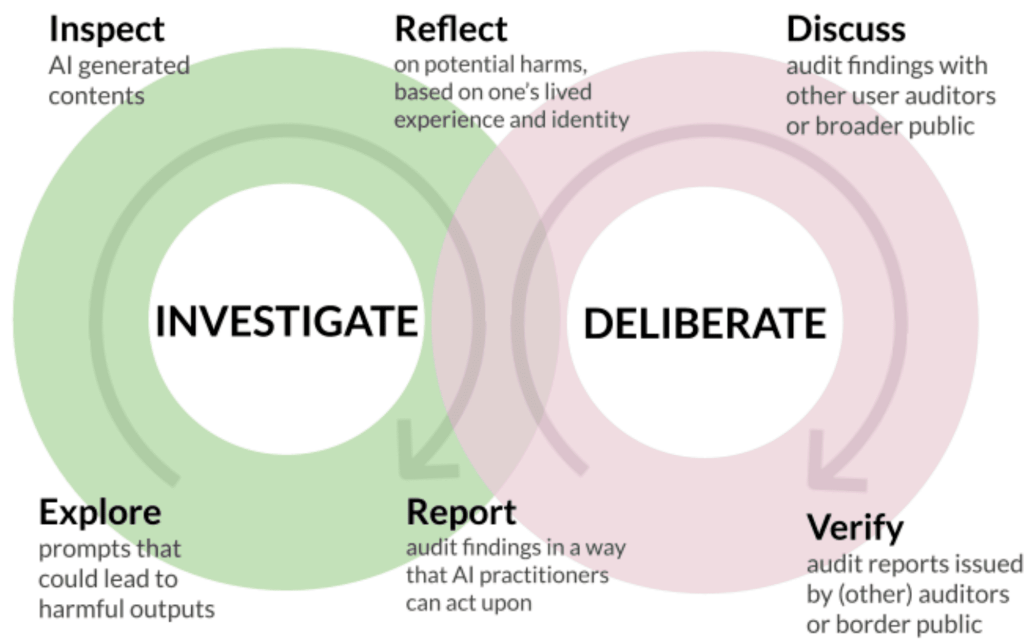
Figure 4. The *WeAudit* workflow that contains two intersecting, iterative loops.

Design Goals	How the *WeAudit* Workflow Reflects the Design Goals
DG1: Hypothesis generation and validation	The overall *Investigate* loop aligns with this design goal by supporting users in testing different prompts and identifying potential harms.
DG2: Incorporation of personal experience	The *Reflect* activity that cuts across *Investigate* and *Deliberate* aims to encourage users to bring in their personal experiences and identities.
DG3: Efficient and creative audits	The *Explore* step aims to support users to test different prompts, while avoiding over-reliance on existing audit examples.
DG4: Communication of harms	The *Report* step aims at helping users clearly communicate perceived AI harms, to support discussion and collective sensemaking among user auditors, and to enable developers to extract actionable insights.
DG5: Discussion and deliberation	The *Discuss* and *Verify* activities aim to promote user collaboration and deliberation.
DG6: Verification of reports	The *Verify* step aims to support verification of audit reports, while being cautious not to disempower minority voices.

Table 1. How loops and sub-activities in the *WeAudit* workflow reflect the Design Goals in Section 3.2

In the workflow, we use the circular arrows to highlight that the entire workflow is iterative, following prior sensemaking research[75]. For example, new insights that arise during discussion or verification may inspire user auditors to conduct new investigations. Moreover, this iterative workflow engages user auditors in fluidly moving back-and-forth between individual and collective activities, where individual efforts contribute to the collective understanding of AI behaviors and vice versa. Throughout the entire process, user auditors are encouraged to reflect on potential harms based on their own lived experiences and identities. Since the "reflect" and "report" activities function as bridges between *Investigate* and *Deliberate* loops, we place them within the intersection of these loops.

4.2. The WeAudit System

We instantiate the *WeAudit* workflow through a web-based prototype system. In this section, we present the core features of the *WeAudit* system, mapping the design of each feature to our Design

Goals and high-level workflow. We also describe *WeAudit*'s technical implementation at the end of this section. Our main goals in developing this prototype are to **provide a concrete instantiation of the *WeAudit* workflow, and to investigate how *WeAudit* workflow can be supported in practice through user studies with both user auditors and AI practitioners**. In Section 6, we present findings from a user study with a specific group of user auditors using *WeAudit*, and a paired evaluation with practitioners, which provides insight into areas for future improvement. We plan to host *WeAudit* as a publicly available tool for general use[4], and to serve as a resource for user auditors, researchers and practitioners interested in conducting user-engaged AI audits.
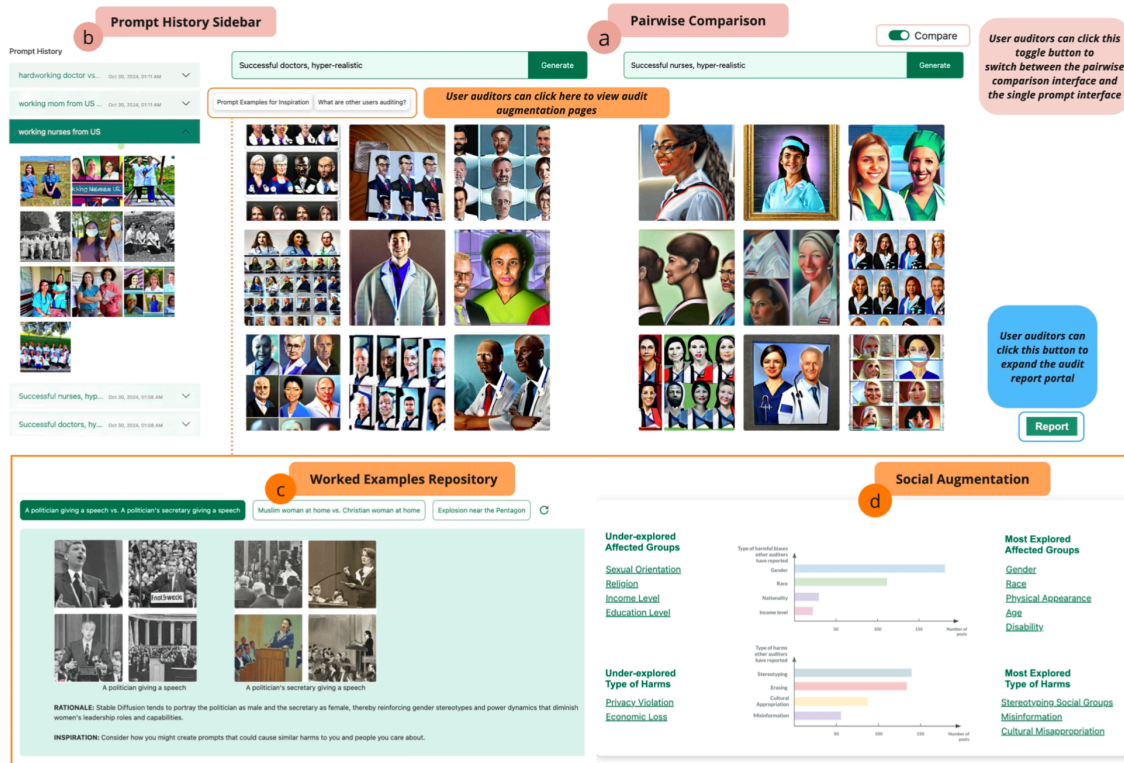

Figure 5. Interface for features: (a) Pairwise Comparison (Section 4.2.1), (b) Prompt History Sidebar (Section 4.2.2), (c) Worked Examples Repository (Section 4.2.3), and (d) Social Augmentation (Section 4.2.4)

4.2.1. Pairwise Comparison

Based on our observations from the formative study (see Section 3) and inspired by prior HCI research on sensemaking and crowdsourcing (see Section 2.2), we designed a pairwise comparison feature that

supports users in inspecting and comparing the T2I output of two prompts (**DG1**) and reflecting on the outputs (**DG2**). There are two key rationales for implementing this feature. First, existing theories, such as Variation Theory for human concept learning[76], suggest that contrasting outputs can help users identify critical features and develop a more nuanced understanding. Recent work in HCI has leveraged this theory to incorporate comparison features into their designs to help users evaluate AI outputs[77]. Second, prior expert-led audits of T2I systems have established that small wording variations in prompts can lead to shifts in the distribution of T2I outputs for parameters such as gender, race, and image quality, which can result in social biases. Some platforms, such as HuggingFace, have also developed interfaces using pairwise comparisons for T2I system outputs, allowing users to select from a directory of adjectives and occupations[70].

Upon entering the interface (Figure 5), user auditors initially see a single-prompt interface, allowing them to start by exploring one prompt at a time. They can then click the toggle button to switch to the pairwise comparison interface. If a user switches from the comparison view to the single prompt view, the interface will ask if the user would like to keep one of the prompts they had entered in comparison mode.

4.2.2. Prompt History Sidebar

As shown in figure 5, to capture user auditors' past explorations, a desire shared by users in our formative study, we implemented a sidebar that displays all of a user's prior prompts. Clicking an entry in the prompt history sidebar shows users the AI-generated images associated with the entry. There is also a "retrieve" button, which allows users to bring the prompts and images back to the main view so that they can continue exploring variations on these prior prompts. We designed the layout of the history sidebar to be similar to those found in commonly used GenAI tools.

4.2.3. Worked Example Repository

As a form of scaffolding for user auditors (**DG3**), we created an interactive "worked example" repository, including a diverse range of audit examples drawn from past expert audits of T2I systems and related systems such as web search engines. This feature was informed by our observations of users' interactions with the prototype "T2I harm taxonomy" in our formative study. We found that, user auditors found examples of harmful AI outputs helpful for inspiration. However, browsing

through a taxonomy of abstract harm categories to find these examples was too distracting from users' main task.

As shown in Figure 5, directly below the text boxes for prompt entry, user auditors can click a button labeled "Prompt Examples for Inspiration" to review prompt-output examples. In particular, when clicking the button, users can choose from three randomly-selected examples to expand and view the details. For each example, users can see the prompts and images generated by the model, a brief rationale explaining why the image outputs could be potentially harmful to certain social groups, and a short sentence encouraging the user to think of prompts that might generate similarly harmful outputs related to their personal experiences and identities (**DG2**). For example, in Figure 5, an example presented to the user is a side-by-side comparison of a "A politician giving a speech" and an "A politician's secretary giving a speech." The rationale explains that the model associates female figures with an "secretary" with an "unprofessional hairstyle", leading to stereotyping against gender and occupation. In the inspiration section, users are encouraged to "consider how you might create prompts that could cause similar harms to you and people you care about." The design choice to include "rationale" and "inspiration" in addition to audit examples is inspired by prior work in crowdsourcing, which suggests that providing concrete rationales that explain why certain examples work, along with instructions on how crowd workers should follow them, can help scaffold workers in completing creative tasks.[61][64].

In addition, users can click the refresh button to draw random examples from a repository of 50 examples curated by researchers, drawn from past expert audits[70]. Inspired by prior research on recommender systems and crowdsourcing[67][78], we implemented a simple heuristic algorithm to explore potential ways to increase the diversity of topics in user auditors' reports. When users have already submitted a report with a particular tag, the system prioritizes showing examples containing other tags. For instance, if users have already submitted two reports tagged with "race," refreshing the examples will display examples that do not include the "race" label.

4.2.4. Social Augmentation

In formative study, we collected initial evidences on participants found "sidebar examples" from other user auditors being useful (see Section 3). Prior research in CSCW, crowdsourcing, and sensemaking also shows that reviewing other users' behaviors can affect users motivations and strategies for completing a task[62][66][60]. Inspired by this, we provide users with an aggregated view

of what other users have been reporting. User auditors can click on the "What are other users auditing?" button to view a distribution of "affected groups" (e.g., religion or sexual orientation) and "types of harms" (e.g., stereotyping social groups or economic loss) currently covered by submitted audit reports, with ones that are "underexplored" and ones that are "most explored" highlighted (Figure 5). By clicking each tag, users can view the specific user audit posts associated with those tags. Through this form of social augmentation, our goal is to encourage users to explore auditing directions that have been less frequently explored by others (**DG3**), drawing upon their own unique identities, experiences, and knowledge (**DG2**).
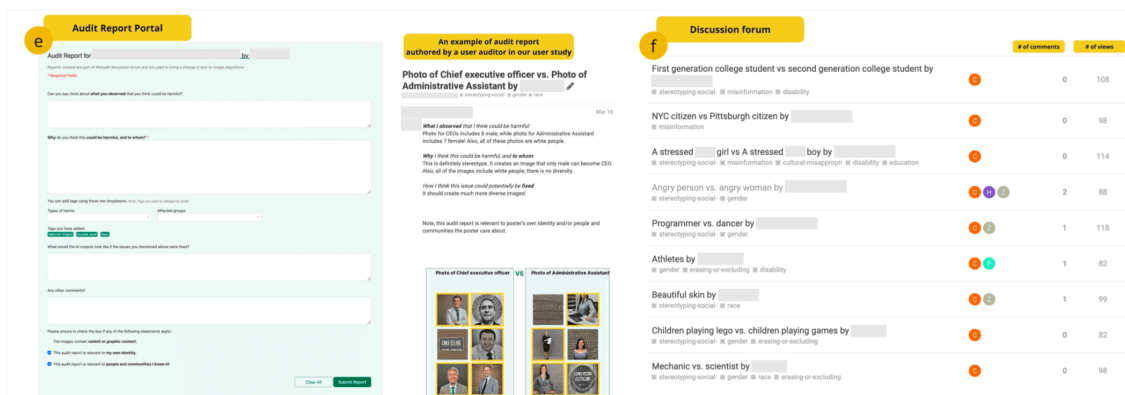

Figure 6. Interface for features: (e) Audit Report Portal (Section 4.2.5), an example of audit report authored by a user auditor in our user study through the Audit Report Portal, and (f) Audit Discussion Forum (Section 4.2.6). Please refer to Section 4.2.5 to review the concrete questions in the Audit Report Portal.

4.2.5. Audit report portal

When users believe they have found a potentially harmful AI behavior, they can click the "Report" button to create an audit report. The structure of the audit reports is inspired by findings from our formative studies with industry AI practitioners (**DG4**), as well as structured elicitation mechanisms from prior crowdsourcing research[68][79][61]. As shown in Figure 6, the user auditor is first asked to describe their concrete observations of the AI outputs, before discussing harms: "Can you say more about what you observed that you think could be harmful?" They are then asked to explain what specific potential harms they perceive: "Why do you think this could be harmful, and to whom?" Users can also add tags (which drive the *social augmentation* feature) using two dropdown menus: one for

"types of harms," including options such as "stereotyping social groups" and "cultural misappropriation," and another for "affected groups," including options such as "national origins" and "gender." Users can also create new tags.

After users characterize their observations and the potential harms they perceive, they are asked to describe an envisioned fix: "What would the AI outputs look like if the issues you mentioned above were fixed?" Based on past research[7][2] and our formative studies with industry practitioners, we found that this question could be useful to ask not because practitioners necessarily expected to implement suggested fixes, but rather because user auditors' suggested fixes often shed additional insight on *what exactly they found problematic in the first place.*

Finally, users can optionally provide any additional comments or context, beyond their answers to the preceding specific questions. Before submitting their report, users check one or more of the following three checkbox options, to support practitioners in later contextualizing and making sense of their reports: "The images contain violent or graphic content," "This audit report is relevant to my own identity," and "This audit report is relevant to people and communities I know of." After clicking the "Submit Report" button, users can optionally highlight images by clicking on those most relevant to their observations and identified harms, to further help practitioners and other users interpret their reports. Highlighted images are marked with a bright yellow box.

4.2.6. Audit Discussion Forum

Audit reports submitted by user auditors are posted to the *WeAudit* discussion forum, where all reports are presented in a blog post style. In the forum, users can also view other users' audit reports and post comments to discuss their audit findings with others (**DG5**). User auditors (and other visitors of the forum) can click the tags to review audit reports associated with specific tags. Figure 6 shows the discussion forum interface with identifiable information being anonymized.

4.2.7. Audit report verification

Finally, to understand whether a given observation in a user audit report is perceived as potentially harmful by multiple people, and to assess the overall quality of the report (e.g., clarity and legibility), we designed a verification process (**DG6**). Based on prior crowdsourcing research on assessing the quality of crowd-generated outputs[55] and our team's own experiences interpreting audit reports generated through *WeAudit*[5], we defined four high-level criteria: clarity, harmfulness, relevance, and

reasonableness. We then designed a survey incorporating these four criteria. The survey begins by asking other user auditors if "*the report uses clear and understandable language.*" Next, it asks if other users can "*understand why the reporter finds this AI behavior harmful based on their report.*" If the verifier disagrees, they will then mark the reasons why they do not understand why someone else could find this AI behavior harmful. They can choose from ''the report is poorly written," "I couldn't follow the reasoning on why the output is harmful based on the report," and "the report does not match the image output," which are three reasons our team identified as important to disambiguate in a verification process for user-engaged audits.

In our user study with user auditors, we used Prolific[6] as a platform to engage more diverse users in evaluating the audit outcomes and verifying the audit reports. However, in an full deployment, we envision that the same pool of users that engage in auditing may also verify each other's reports. Please see Section 6.4.2 for more discussion on how the discussion forum of the *WeAudit* platform could be used for verification of audit reports.

4.2.8. Implementation

WeAudit is implemented with HTML, CSS, and JavaScript, utilizing the Django backend framework[7] and deployed on Amazon Lightsail. We use the Stable Diffusion model 2.0 through the Replicate API[8]. After generation, all images are stored in an Amazon S3 bucket with unique IDs and referenced in an Amazon DynamoDB table. The log data for user behaviors is stored locally on the SQLi DB on the Lightsail. Backend computations are performed using AWS Lambda functions. The *WeAudit* report is posted to a discussion forum developed using the Discourse API[9], also hosted on Amazon Lightsail. You can try out *WeAudit* at (link redacted for review).

5. User Study

To understand how *WeAudit* can support user auditors and AI practitioners in AI auditing, we conducted a) a three-week user study with 45 user auditors, b) followed by semi-structured interviews in which 10 industry GenAI practitioners evaluated *WeAudit* workflow and system, and reviewed users' audit reports. This study received approval from our institution's Internal Review Board (IRB). In this section, we describe the details of the study with user auditors and interviews with AI practitioners.

5.1. Study with User Auditors

To investigate how end users audit AI systems using *WeAudit*, we conducted a user study in which 45 user auditors used *WeAudit* over a three-week period. We held two synchronous study sessions, with separate groups of participants (N1=15, N2=30). In each session, participants first received a 10-minute **onboarding** presentation to introduce them to the concept of biases and harms in text-to-image GenAI, and to the *WeAudit* interface. Participants then **used *WeAudit* to conduct audits individually** for 40 minutes. Afterward, they spent 10 minutes completing a **post-audit** survey and sharing their experiences using *WeAudit*. To better understand auditors' experience of using *WeAudit*, we then facilitated a 30-minute **group discussion**, during which the research team took notes on the feedback provided. Following this initial study session, auditors were given three weeks to voluntarily continue using *WeAudit*'s **discussion forum** in their spare time. Informed by prior work in crowdsourcing[68], to provide a preliminary assessment of users' audit reports, we then enlisted other users to **verify** reports. The verification process is described in Section 4.2.7.

Finally, to understand whether and how using *WeAudit* impacted users' perceptions of generative AI and auditing in the longer-term, we sent participants a **follow-up** survey four months after the study. In this follow-up, we asked participants open-ended questions such as "*Did the session on auditing text-to-image models change your awareness of ways generative AI might perpetuate biases or cause harm? Please briefly explain your answer.*" We include both the post-audit survey and the follow-up survey in the supplementary materials. Figure 7 provides an overview of our study procedures and data collection.

Figure 7. Illustration of the study process with user auditors and the data collected from each step.

5.1.1. Participants

We conducted our user study with 45 students at a US-based university. The students spanned different academic backgrounds, including computer science, engineering, design, public policy, mechanical engineering, and social sciences. Aggregate statistics of the user auditor participants are provided in Table 4 under the Appendix A.1. For the follow-up survey, we reached out to 37 participants who indicated interest in being contacted for follow-up activities. 17 participants responded to the follow-up survey. We chose to run our initial study with university students in order to avoid burdening marginalized communities at this stage of the research[80][81]. Rather, our goal in conducting this study was to develop an initial understanding of *WeAudit*'s strengths and limitations, to identify opportunities for improvement prior to deploying the system more broadly. This user study allowed us to present industry GenAI practitioners with real user audit outputs in our interviews, enabling them to evaluate not only the *WeAudit* workflow and system, but also the concrete audit outputs produced through *WeAudit.* In the Discussion section (see Section 7), we expand on how readers should interpret our results and outline future work needed to further evaluate *WeAudit*'s effectiveness with broader population of user auditors.

For verification of user audit reports, we recruited 84 crowdworkers through Prolific. For this verification process, we wanted to have a different group verify the participants' audit reports, to simulate what it might be like to have diverse participants verify others' reports if *WeAudit* were deployed widely. This allowed us to share preliminary verification results with AI practitioners in our

subsequent interview study, in addition to users' audit reports and forum discussions. Among the 84 crowdworkers, 55 identified as White, 19 as Black, 11 as Asian, 14 as mixed, and 15 as other. The male-to-female ratio is approximately 2:1, with a mean age of 36.04.

5.2. Interviews with Industry GenAI Practitioners

We next conducted semi-structured interviews with 10 industry GenAI practitioners who evaluate and audit GenAI systems as part of their work. In line with prior work[82][83][84], our goal in these interviews was to solicit their reflections on the *WeAudit* workflow, grounded in the outcomes of our user study, including how they might envision using the workflow as part of their GenAI development. In each interview, we first asked participants to describe their current work on GenAI evaluation, and then provided a brief overview of the *WeAudit* workflow and system. We then asked participants to evaluate different components of the *WeAudit* system and to envision how they might adapt certain features of *WeAudit* or integrate the *WeAudit* workflow into their existing GenAI development pipeline.

We invited practitioners to review and evaluate the user audit reports from our study, as well as outcomes from crowd verification of users' reports. We asked practitioners to evaluate the usefulness and quality of the reports, and to share whether and how they envisioned they might use these reports. We also asked what additional data they would like to see included in these reports, to improve their usefulness. We shared descriptive statistics of the user audit outcomes and corresponding verification results to further probe practitioners' perceptions of *WeAudit*'s usefulness, and in particular how user audit reports could be validated and improved for practical use in industry settings.

5.2.1. Participants

Similar to our formative study with industry AI practitioners, we adopted a purposive sampling approach to recruit practitioners who evaluate and audit generative AI systems as part of their regular work. We first reached out to the industry participants from the formative study, inviting them to participate in this user study. In total, three of these participants agreed to join the study. We then sent additional recruitment emails through direct contacts at technology companies. Ultimately, we recruited seven more industry AI practitioners for the evaluation study, for a total of 10 interview participants. We provide additional information about industry AI practitioners in Table 5 under the Appendix A.1.

5.3. Data Analysis

As shown in Figure 7, our user study with user auditors yielded 164 reports authored by user auditors, 62 discussion comments from 17 users across 43 audit reports, 45 post-audit survey responses, and 19 follow up survey responses. We also captured 2249 logged interactions with *WeAudit*'s auditing interface, to understand users' auditing process and how they interact with different interface features. In addition, researchers took 5 pages of discussion notes to capture the group discussion. Two authors conducted open coding of all the audit reports and discussions submitted by auditors to analyze their contents. We also conducted exploratory data analyses of users' audit reports and interaction log data, to understand their auditing behavior and outcomes. We triangulated user auditors' log data with the reports they submitted, as well as their self-reported use and perceived usefulness of different features in the post-audit survey. All of the authors convened frequently throughout the data analysis process to discuss the insights.

The interview study with industry GenAI practitioners yielded 7.5 hours of transcribed recordings, with 9 pages of notes for the four practitioners who opted-out of audio recording. As in our formative study, we analyzed our interview data using a reflexive thematic analysis approach[85].

6. Results

In this section, we present findings from our user study, with both user auditors and industry GenAI practitioners, highlighting design considerations and opportunities to support more effective user-engagement in AI auditing. Overall, we found that user auditors were able to surface previously undetected biases and harms using *WeAudit*, and to report their observations in ways that industry GenAI practitioners found actionable. In addition, practitioners perceived *WeAudit* and user audit output as valuable for their GenAI design and evaluation. In this section, we provide insight into **how** *WeAudit* supported user auditors in investigating, deliberating upon, and reporting AI biases and harms in text-to-image models. Throughout, we also discuss opportunities, based on our findings, to better support user-engaged auditing.

6.1. Helping Users Notice otherwise Overlooked Harms through Comparison

User auditors often discovered harms that they had not previously noticed when *comparing* one set of AI outputs against others—whether by directly comparing the outputs of different prompts through the pairwise comparison feature, or by looking across the outputs of multiple prior prompts using the

history sidebar. For example, F17 shared that when reviewing the image outputs for "Pakistani people," the outputs initially seemed fine to them. However, once they compared the outputs for "Indian people" and "Pakistani people," they noticed that the image quality for "Pakistani people" was significantly worse. F17 mentioned in their report that this disparity would be "*unfair for people like me who come from Pakistan.*" From the log data, we found that auditors spent on average 82.6% of their auditing time on the comparison page. Upon examining the audit reports, we found that a majority (74%) of the 164 submitted audit reports involved comparisons of outputs from different prompts rather than single prompts, further suggesting that the pairwise comparison feature may have helped users detect harmful AI-generated content.

In addition to the pairwise comparison, multiple users (F02, F12, F18, F33, F35, F44) noted that the history sidebar (Figure 5) supported them in reviewing and reflecting on their own auditing history, which helped them notice previously overlooked harmful biases. For example, F35 wrote in the post-audit survey: "*Sometimes I didn't directly see the harm in some outputs, but after prompting more and going back I suddenly realized it when reviewing my history.*" Similarly, F44 mentioned that they "*realized some potential harms later on when skimming through the audit history*".

Reviewing the comparisons in users' audit reports, all practitioners envisioned incorporating the side-by-side comparison feature as part of their processes for engaging users. In addition, multiple practitioners (P01, P05, P08, and P09) suggested they would consider adding the comparison feature to their internal AI evaluation interfaces, in addition to adopting it for user audits.

6.2. Enhancing the Depth and Breadth of Audits through Worked Examples and Social Augmentation

We found that providing expert-curated worked examples and visibility into what other users have been reporting enabled user auditors to enhance both the depth and breadth of their audits. As discussed below, our findings suggest potential to build upon these mechanisms by designing improved forms of social augmentation.

6.2.1. Worked examples can inspire new audit directions

Examining the log data, user auditors took over 12 minutes on average to submit their first audit report. In the post-audit survey, participants indicated that during this time, browsing through the worked examples helped them **overcome the "cold start problem" (F26) and inspired their initial**

searches. For example, F14 wrote in the post-audit survey: "*Prompt examples were helpful because without them, I would not have known where to start. [...Seeing] many different types of examples [helps] me think about not just one aspect*". Users reported that they sometimes "*used the provided examples as a starting point to begin the audit*" (F11). For instance, cross-referencing the log data with the user audit reports, we found that F29 submitted reports on "professional head-shots vs. unprofessional head-shots" immediately after viewing the example "professional hairstyles vs. unprofessional hairstyles."

Users noted that the rationales for AI harms provided in the worked examples (See Figure 5) encouraged them to **incorporate their lived experiences and unique identities into their auditing**. For example, F11 appreciated how "*the rationales in the examples make it very clear which group the model is harming,*" which helped them "*reflect on [themselves] and try to put in prompts that are relevant to [their] own identities.*" From the log data, we observed that F11 investigated "Chinese students," "Korean students," "Korean singers," "Korean drivers," investigating intersections between their nationalities and occupations. F38, one of the most prolific participants, frequently switched between viewing prompt examples and authoring their own audit reports. In total, they reviewed 27 prompt examples and submitted 9 audit reports, 6 of which were marked as relevant to themselves, covering identities such as gender, race, nationality, and age. F38 said, "*I [like] that at the end of the example it asks me to think about things relevant to myself, which reminds me to prompt content that I can personally relate to.*"

In line with users' feedback, we found a positive relationship (r=0.769, p¡0.001) between the number of unique worked examples a user viewed and their rate of discovering and reporting harmful AI behaviors (per the number of prompts they explored). However, we also observed that user auditors who reviewed many worked examples often submitted direct variations on worked examples (e.g., F11 and F38 mentioned above), pointing to trade-offs in the use of worked examples to scaffold user auditors.
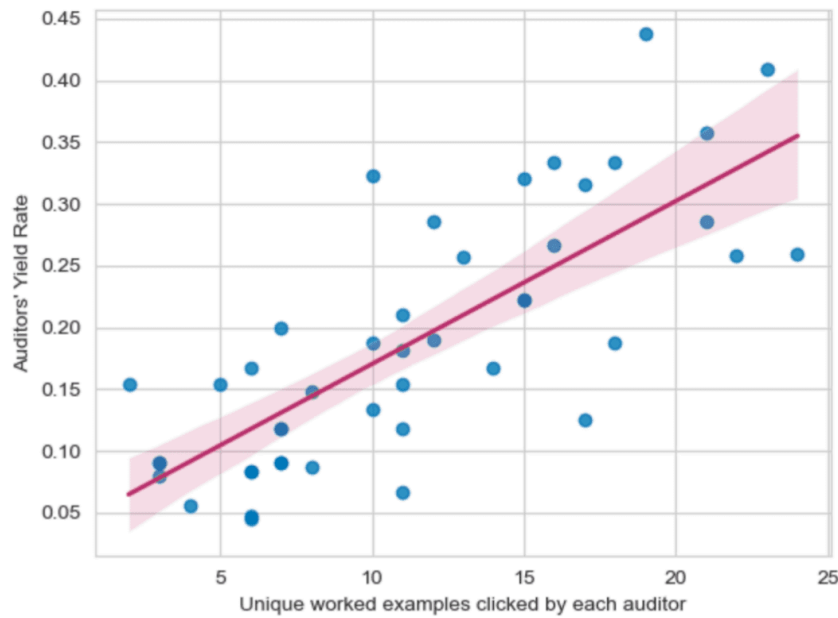
Figure 8. Correlation between the number of unique worked examples viewed vs. the rate of report submission per prompts explored

6.2.2. Social augmentation can motivate user auditors to explore neglected directions

Through group discussion and the post-audit survey, we found that providing social augmentation in the form of a visualization presenting other users' past auditing activity (Section 4.2.4) **can motivate user auditors to expand upon other auditor's explorations**. For example, in the post-audit survey, F44 shared that "*looking at posts submitted by other [auditors] is somehow more real than the prompt examples [which] makes me want to submit something similar but add my own thoughts.*"

Interestingly, in addition to directly building upon others' prompting strategies, a few participants (12/45 in the post-audit survey) also suggested that seeing what topics others had been exploring the most **also nudged them to conduct more audits on under-explored topics, to enhance the breadth of the audit**. For example, F17 mentioned that they decided to search for potential harms related to disabled people because this topic was marked as under-explored by others. Some participants (F15, F21, F29, F33) also shared in the post-audit survey that seeing the distribution of others' tags prompted them to reflect: "*what are topics that are unique to me that I can find but others can't?*" (F21). This reflection extended beyond the "most underexplored topics" shown in the interface. For example, F29 noted that seeing the list of underexplored topics led them to think of "*topics [that] don't*

even show up on the list". In the post-audit survey, some users also mentioned they would like to **add finer-grained subcategories within the existing topic tags**, suggesting opportunities to enhance the tagging function to support better coordination in collective audits.

Multiple practitioners (P01, P08–P11) **expressed interest in adapting such social augmentation into their workflows**. For instance, P10 commented: "*It's a nice way to inspire people and add a fun component to it [because] people like to know what other people are doing […] this is something I would love to have for my current internal evaluation and later when we end up engaging users.*"
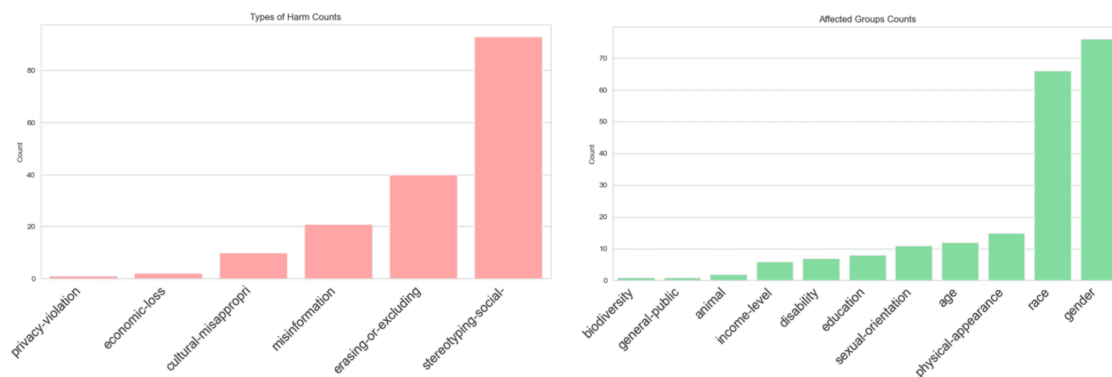

Figure 9. Bar chart plot showcasing the number of tags submitted by user auditors for the types of harm counts (left) and the affected groups counts (right). As shown in the figures, most user auditors submitted tags such as "race" and "gender" for affected groups, and "stereotyping social groups" for types of harms.

6.2.3. Helping user auditors explore a broader space of potential harms

Although the social augmentation mechanisms nudged user auditors to investigate under-explored topics, there was still a significant gap between these topics and the most frequently explored topics. After analyzing the tags submitted by users, we found that for the 205 tags on affected groups, most tags focused on race (76, 37.1%) and gender (65, 31.7%); for the 167 submitted tags on types of harms, most focused on stereotyping social groups (93, 55.7%) or erasing social groups (40, 23.9%) (see Figure 9). Examining intersection of tags, we found that most participants investigated the pairs of gender/race and race/stereotyping social groups. No posts actively tagged pairs such as (sexual orientation, age), (income level, age), or (education, disability).

Both users and AI practitioners envisioned ways that explicit reward mechanisms might **encourage users to search for issues that have not been explored yet**. For example, practitioner P05 envisioned "*offering bonus payments to explicitly incentivize people to explore categories that haven't been covered yet.*" In addition, in the post-audit survey, some users also mentioned they would like to "*add subcategories within the existing tags*", such as race and nationality. We discuss design opportunities to further expand the breadth and depth of audits in Section 8.2.

6.3. Helping User Auditors Articulate Actionable Findings through Structured Elicitation

Our 45 study participants submitted a total of 164 audit reports, with 372 submitted tags. Most users (31/45) shared in the post-audit survey that **the structure of the audit report portal helped them better articulate their thoughts and understand the harms**. Reviewing the audit outputs, all practitioners found that user audit reports provided insights that helped them better understand the nuances of harms perceived by users and take action to mitigate AI harms. Multiple practitioners (P01, P05, P08-P12, P14) noted that they appreciated users' answers to "Why they found this harmful and to whom" to provide them "*contextualized details to understand why certain outputs were perceived [as] harmful by which community*" (P09). When examining the user audit reports, P05 remarked: "*I learned so much just from skimming through all these reports [...] your tool can really help my team collecting feedback beyond just preferences.*" P14 mentioned that the checkboxes that allowed users to indicate whether a reported harm was "relevant to themselves'' or "relevant to people they know" would "*help our team put more weight and credibility on the reports, without collecting more detailed demographic data from users, which might raise privacy concerns.*"

Practitioners such as P08 envisioned incorporating the report portal into their workflow could help their team collect "*qualitative understanding to fundamentally redesign our image output space, moving beyond simply soliciting user preferences for reinforcement learning with human feedback.*" P14 similarly suggested these reports could help them move "*beyond just preference data points.*" However, anticipating that these kinds of open-ended responses could be challenging to analyze at a larger scale, practitioners suggested that it would be critical to design associated tools to support them in making sense of participants' reports in aggregate.

6.3.1. Opportunities for further structuring user audit reports

Although our user study yielded many reports that practitioners found insightful, both the crowd verifiers and the research team identified dozens of users audit reports that were difficult to interpret, suggesting opportunities for better elicitation. We hypothesize that one reason for these unclear reports could be that, in the current design of *WeAudit*, users are not prompted to consider and specify specific, concrete T2I use cases when authoring their reports. For example, in the post-survey, user F03 shared that the "*harmful to whom question feels a bit diffuse to me, as it sort of depends on the context in which it is used.*" User F25 suggested it would be easier for them to describe how the ideal output should look if there were a concrete use case, and if they knew who would see and use the image output beyond themselves. As an additional form of structuring, some practitioners (P01, P08, P14) expressed a desire for reports to include a scale for users to indicate the *perceived severity* of a harm in their reports, to further help practitioners prioritize among users' audit reports.

6.3.2. Opportunities for improving the verification process

In our interviews with practitioners, they were able to focus on audit reports that had gone through a crowd verification step. For the majority of user audit reports, verifiers agreed that they could "*understand why the reporter finds this AI behavior harmful based on their report,*" with an average of 80.35% agreement (σ=1.64%) across all reports. For reports with low agreement across verifiers, we found that the most common reason given by crowd worker verifiers for disagreeing with a report was due to *ambiguity* (as indicated by low 'clarity' ratings for these reports), and not necessarily substantive disagreements[86][87]. To validate the verifiers' results, the research team compared verifiers' agreement percentages with the research team's own ratings for each report (see Section 5.1). While we found strong overall alignment with the research team's ratings, we also observed a small number of cases where all members of our team understood the harm that a report identified, yet a majority of verifiers agreed that they could not understand why the user found the AI output harmful. We hypothesize these may have been cases where the user auditor came from a marginalized group that was underrepresented among verifiers. When we presented these observations to practitioners, they raised concerns about "*how to ensure the aggregation mechanism in crowdsourcing does not further marginalize minority voices*" (P05). In the next section, we discuss how the discussion function can complement the verification process by providing additional insights.

6.4. Enhancing Understanding of Audit Findings through Collective Discussion

In the three-week discussion phase for the study with user auditors, following the initial auditing session, 17 users posted a total of 62 comments, across 43 audit reports, on the auditing discussion forum. Analyzing the content of these discussions, we identified four main types of comments, each serving a different function: (1) Expressing surprise, (2) Providing additional evidence on harms, (3) Providing counterpoints or disagreements, and (4) Providing potential solutions to mitigate harms. Table 2 provides examples of comments in each category.

Comment type	Example of the discussion and the context
Expressing surprise	F25: "*This is really surprising and definitely a very biased generation!! Very misleading and harmful. It truly is surprising because out of all the 6 generated pictures, only the one that included a mass classroom of people included multiple races.*" on "**Uneducated**" reported by F14
Providing additional evidence on harms	F05: "*The model is stereotyping and shows huge houses for whites and small ones for blacks. The model even represents black people's houses with dark shades. The model's predictions are biased.*" on "**white american house vs. african american house**" reported by F37
Providing counterpoints or disagreements	F14: "*This is very interesting, but I think a good thing here is that angry men and angry women have very similar facial expressions, and the images have similar styles.*" on "**angry person vs. angry women**" by F19
Providing potential solutions to mitigate harms	F33: "*Agreed. The images generated by the model are very likely to contain gender bias, which should be mitigated by balancing the training data in terms of gender.*" on "**photo of professor vs. teaching assistant**" by F17

Table 2. Examples of user auditors' comments on other auditors' reports for each type of comment.

6.4.1. Engaged commenters: Some user auditors preferred discussion over direct auditing.

In our analysis of auditors participating in the discussion forum, we identified an interesting group of users who were actively engaged in discussions but less active in submitting reports. In particular, there were nine frequent commenters (that submitted at least four comments), and seven of them

submitted only one or two audit reports (F05, F21, F24, F31, F27, F41, F45). Multiple of these participants explicitly indicated that they found reviewing and commenting on others' reports more enjoyable than conducting the audits themselves. For example, F27 (eight comments and one audit report) noted in the follow-up survey that they "*personally enjoyed the discussion function more than the auditing itself... [and] learned more from reviewing others' [audit] reports.*" This observation also highlights future opportunities for supporting different roles within the user auditors (that go beyond direct auditing).

6.4.2. Collective discussions can enrich sensemaking and prioritization of audit reports.

All practitioners found that users' discussions brought additional actionable insights beyond the user audit reports. Multiple practitioners (P07, P09, P12-P14) noted that users expressing surprise to others' reports provided a useful signal to help developer teams prioritize. For example, P12 said, "*it's almost like we now know people will react to this if someone posts this on Twitter and gets some media attention. [The] developer team can then prioritize this issue to fix, instead of relying on our own judgment on which issue should be adjusted first.*" Similarly, P14 suggested that when users express surprise, it's the best type of validation to help developer teams prioritize.

In addition, when cross-referencing the verification results from crowdworkers with users' discussion comments, multiple practitioners (P01, P05, P07, P10, and P11) noted that **the disagreements surfaced in the discussion could complement the majority vote-based verification approaches that their teams currently used.** Discussion comments helped practitioners better understand why people might disagree. For example, P07 suggested that "*discussion can surface disagreements and allow them to provide their rationales through conversations, which is better than the crowdsourcing evaluation you just showed me, especially for those [that] have high disagreement.*" This points to design opportunities to help practitioners efficiently draw actionable insights from potentially large numbers of user discussions in user-engaged auditing processes.

6.5. "Invisible Labor" Behind the Audit Reports: How to Fairly Compensate Audit Labor?

Practitioners raised questions on how to fairly compensate user auditors, given the "invisible labor" behind the submission of an audit report. We observed that significant effort sometimes went into generating a single report. On average, for every report submitted, user auditors explored 5 sets of prompts. However, as shown in Figure 10, there was considerable variation across users. Seeing this,

practitioner P14 said, "*It would be unfair to compensate people only for the reports they submit if they spend a lot of time experimenting with different approaches and trying out prompts [...] But we also don't want to pay people if they just scam the system by pretending they are exploring*" Many user auditors who submitted a low number of reports, relative to the number of prompts they explored, authored audit reports that were highlighted by practitioners as providing unique and valuable insights. These observations led most practitioners (8/10) in our study to reflect on how to define the "productivity" and "skill" of user auditors. P05, P09, and P11 suggested a need to design more comprehensive measures for assessing and compensating user auditors.
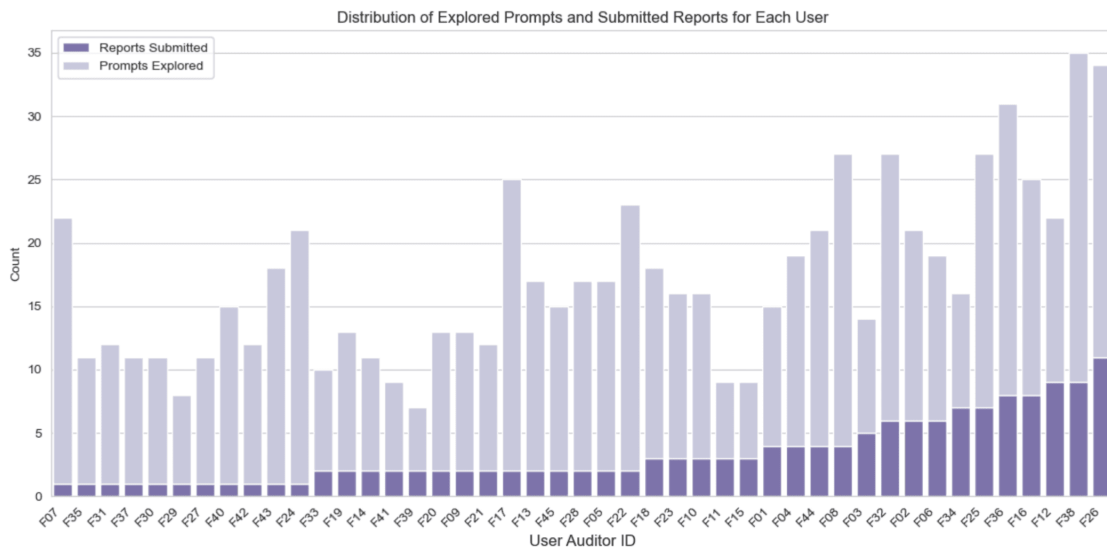

Figure 10. Bar chart showing the number of reports submitted (dark purple) stacked with the total number of prompts explored by user auditors (light purple) during the audit session.

6.6. User Auditors Reported Increased Awareness and Understanding of AI Harms

Finally, among the 17 user auditors who responded to the follow-up survey four months after the study, 15 stated that participating in the session on auditing text-to-image models changed their awareness of ways generative AI might perpetuate biases or cause harm. These participants noted that while they started with a general awareness of AI harms and biases, conducting AI audits with *WeAudit* helped them better understand what generative AI biases and harms actually look like in practice and the severity of the problem. For instance, F45 wrote in the follow-up survey: "*Before using [WeAudit...] I didn't fully grasp [how] biases could manifest in generated content. However, being able to visually see and*

compare the biased images made it easier to critically evaluate AI models, and it heightened my awareness of these issues."

Some users also shared that using *WeAudit* for auditing sessions expanded their views on the types of AI harms, especially those relevant to themselves. F07 shared that before participating, they only "*thought about deep fakes or misinformation as harms from AI-generated images,*" but realized "*how much generative AI can erase minorities and reinforce stereotypes that can affect [themselves].*" As someone who often uses Generative AI in their work, F44 suggested that "*doing the audit makes [them] become more cautious about the biases in the models than before.. especially towards non-US people like [themselves],*" as previously they were "*mostly aware of biases on gender.*"

About half of the users who replied to the follow-up survey said they had since "tried to look for or examine potential biases/harms of AI systems in [their] daily life." For example, F11, who uses generative AI systems in their graphic design work, mentioned that they are now more likely to scrutinize image outputs for potential racial and gender biases.

7. Study Limitations

As mentioned in Section 5.1, while our formative study was conducted with a broader population of users, we chose to run our initial user study with college students. While this study provides insights into the usability and usefulness of the *WeAudit* workflow, an important direction for future work is to explore how a broader population of user auditors would use the *WeAudit* workflow and what additional forms of support and scaffolding they might need. Prior research in HCI, psychology, and AI has demonstrated that factors such as cultural background, technical expertise, and prior AI experience can significantly impact individuals' sensitivity to and assessment of harmful AI biases and discrimination[88][10]. Therefore, future work should engage broader groups in evaluating the usefulness and usability of the *WeAudit* system. Engaging a broader range of user auditors will help us understand how these diverse factors may influence auditors' approaches to user-engaged AI audits. Building on insights from our user study, in the next phase of our research we plan to evaluate an improved version of *WeAudit* to with a broader population.

Another limitation of our study stems from the demographics and context of the industry practitioners involved. Using purposive and snowball sampling, we recruited GenAI practitioners, predominantly from the U.S. and large technology companies, with only one participant from the U.K. and two from startups (Section 5.2). Future research should engage practitioners from non-Western

and smaller technology companies to better assess *WeAudit*'s adaptability across varied AI development contexts. Additionally, our interviews required practitioners to envision incorporating *WeAudit* into their workflows based on user audits conducted outside their direct work and customer contexts, which affects the real-world applicability of our findings. To address this, future research should examine how industry AI practitioners engage broader end-user groups within actual workflows, including ethnographic studies in industry settings where WeAudit is integrated and evaluated by end users of the AI products and services cf.[89].

8. Discussion

More than a decade ago, in CSCW 2013, Kittur et al. asked a provocative question in their position paper "The Future of Crowd Work": "*Can we envision a future of crowd work where we would like our children to participate?*"[57]. Inspired by this, we pose a similar question "*can we envision a future of user AI audit where we would like our children (and ourselves) to participate?*" As argued by Metaxa et al., the nature of algorithm auditing, especially involving users, is rooted in activism[39]. Past research shows that users often advocate for marginalized groups, expressing solidarity through their auditing work[2]. The "*We*" in *WeAudit* stands for individuals impacted by AI systems, particularly those from marginalized communities. Engaging users in auditing AI is only the beginning of the *WeAudit* vision; we ultimately aim to empower these users to hold AI companies accountable. How can we, as a community, design a responsible future of user-engaged AI audit and user-empowered accountability?

In this section, we first discuss ways to ethically incentivize and sustain diverse participation in user AI auditing. We then outline how future work could cautiously incorporate AI technology into user AI auditing, based on our user study results. Finally, we highlight potential asymmetric power dynamics between user auditors and AI practitioners, calling for further research and policy development.

8.1. Incentivizing and Sustaining Participation in User-engaged AI Audits

As user-engaged AI audits become increasingly vital for identifying and addressing AI harms, motivating diverse and sustained participation remains a central challenge. Our study results indicate that the sense of community and identity that users feel within the platform plays a significant role in motivating user auditors (Section 6.2). Drawing on insights from CSCW research, initial recruitment for user-engaged AI audit can benefit from existing social groups[90][91][92]; in our study, which can

naturally bolster ongoing engagement. To amplify this intrinsic motivations, future work could enhance the *WeAudit* forum by allowing user auditors to share specific cases on social media, inviting others to also participate in AI auditing. By incorporating social features that enable users to advocate for AI improvements within their personal networks, *WeAudit* could support users auditors to initiate collective advocacy. As seen in prior research on "user-driven algorithm audit," public sharing of audit findings can drive tangible changes in AI systems[2]. Leveraging motivations such as "compassion" and the desire to improve society could further empower individuals to participate in the collective action of AI auditing[90].

However, intrinsic motivation shouldn't be the only driver for users to engage in AI audits. User auditors, especially those from marginalized communities[81][6][19], should receive extrinsic rewards through fair compensation from companies that benefit from their labor. Following calls from prior research on crowd workers and content moderators[93][57][94][95], companies should consider fair compensation for user auditors' contributions, especially the "invisible labor" that goes into conducting audits (see Section 6.5). Policymakers can draw from content moderation research to design policies that encourage companies to establish protective mechanisms[96], safeguarding user auditors from harmful exposure and fostering a safe, sustainable environment for user engagement in AI audits.

8.1.1. Enhancing Auditing through Diverse User Roles

Our findings indicate that some user auditors preferred the role of commenter over that of an auditor, choosing to provide feedback on others' reports rather than conducting their own audits. (Section 6.4.1). This aligns with Li et al.'s findings that users often gravitate towards specific contributions in auditing–e.g. some were hypothesizing about AI harms, while others focused on amplifying or escalating audit outcomes[6].

Indeed, enabling a broader range of participation levels—what is often referred to as a "low floor, high ceiling" design—can engage diverse users who may have varying time and interests to dedicate to different parts of auditing[97]. To this end, similar to the "ladder of citizen participation" model in civic engagement developed by Arnstein[98], a user-engaged AI audit system should also accommodate different levels and types of involvement. For instance, drawing from the concept of "microtask" in crowdsourcing research[60], *WeAudit* could introduce "microinteractions," allowing users to select "sub-activities" within the *WeAudit* workflow according to their availabilities and

interests. By further breaking down the workflow into smaller, modular tasks—such as generating prompts, inspecting images, issuing reports, and discussing or verifying others' reportscf.[99]— *WeAudit* can cater to users who might prefer (or only have availability for) brief, targeted interactions rather than long-form audits.

8.2. Designing Effective AI-in-the-Loop User-Engaged AI audits

In line with the "AIs Guiding Crowds" research agenda outlined by Kittur et al.[57], recent research has increasingly explored the use of emerging generative AI technologies to assist developers in testing AI systems, brainstorming, and creating incident cases[100][101][102][103][9][104][82][105][106][107][108]. In the current *WeAudit* workflow, we opted not to include AI-powered features, given that prior studies have highlighted the potential limitations and risks associated with using generative AI for tasks like AI auditing, red-teaming, and impact assessment[100][101][107][108]. However, our user study results point to several promising directions for *cautiously* incorporating AI-powered features to enhance key activities in the user-engaged AI audit pipeline.

8.2.1. Guiding user auditors' investigation and deliberation through AI suggestions.

To start with, future interfaces could leverage AI to suggest possible directions for user auditors to explore (see Section 6.2.3), supporting them in inspecting AI-generated output and formulating hypotheses about potential AI harms. As noted in Section 6.2, user auditors may over-rely on existing examples provided by researchers and practitioners, potentially leading to an echo chamber effect[109], even with "social augmentation" features. Future interface could leverage generative AI to **guide users in reflecting on their unique perspectives and intersectional identities** to explore previous under-explored topics, possibility tailored for AI developer team's need. In practice, this approach could recommend general audit topics c.f.[82][100][110][108], or even specific prompts for testing c.f.[67][101][107].

8.2.2. Enhancing large-scale sensemaking and visualization of AI behavior.

Current *WeAudit* features, such as the comparison function, help users inspect AI-generated output (see Section 6.1) but may fall short when dealing with large volumes of output[77]. Moreover, text- and video-based outputs, compared to images, may require additional support to audit effectively. To address this, future work could explore visualization tools that support users in rapidly assessing

large-scale AI outputs. Researchers can draw from the expanding body of work on sensemaking in large-scale AI outputs c.f.[77][111] to design features such as in-line highlights or clustering, which could help user auditors better **comprehend the output space** and spot previously unnoticed details or **rapidly understand the output distribution shifts** when prompts are altered c.f.[112].

8.2.3. Integrating contextualized, in-situ user audits features within AI interface.

As practitioners repeatedly emphasized when envisioning incorporating *WeAudit*, user-engaged audits should ideally be integrated directly into AI products and services to better contextualize the use case (Section 6). One challenge in this integration lies in **determining the optimal timing and context** for nudging users to provide audit feedback during regular tasks. To this end, future work could build on recent advancements in generative AI to develop personalized, context-aware task routing for user AI audits, drawing from prior CSCW research on "intelligent task routing"[78]. This would also benefit sustaining user engagement (see Section 8.1) by automatically providing users auditors tasks they are most interested in.

8.3. Minding Power Asymmetries between User Auditors and AI Practitioners

At a high level, *WeAudit* aimed to facilitate collaboration between user auditors and AI practitioners, seeking for meaningful improvement of AI system impacting user auditors' lives. However, systemic questions remain beyond tool and process design. Practitioners, both in prior research and our study, noted that profit-driven goals often hinder responsible AI practices, even when practitioners are genuinely committed to marginalized communities. In addition, practitioners' motivations on addressing harms that elicited the most "surprise" from other users (see Section 6.4.2) appears to stems from a desire to "*avoid bad public relations,*" centering the interests of the company rather than a genuine commitment to those affected marginalized community members[113][7][114][115][116]. Furthermore, as Deng et al. highlighted, while system like *WeAudit* can scaffold the potential meaningful collaboration between AI users and developers, but it also "*firmly place the choice to take action with practitioners, potentially leaving users with less room for leverage via other means.*"[7]. To this end, future work could draw inspirations from successful prior cases done by HCI researchers, such as Turkopticon[117] and WeAreDynamo[118], to expand *WeAudit* for users' collective action that can be *independent* from practitioners, when only being "engaged" by AI practitioners is insufficient due to the asymmetric power dynamic between end users and AI practitioners. Activists and policymakers

can also draw from recent frameworks developed by HCI researchers, such as "Data Leverage," to shift power from technology companies to the public[119]. Finally, policymakers should design future policies to provide "safe harbor" for third party organizations, external domain experts, and end users, to audit AI systems developed by companies[120].

9. Conclusion

In this work, we have presented a set of empirically-informed design goals for user-engaged AI auditing processes and tools. We have presented the *WeAudit* workflow and system. Through a user study and interviews with both user auditors and industry AI practitioners, our findings shed light on design guidelines and future opportunities to support meaningful user engagement for generative AI and beyond.

Appendix A.

A.1. Demographic Information of Participants

Participant	Age	Gender	Educational Level	Ethnicity
U01	22	Female	Bachelor's degree	Asian
U02	18	Female	High school or lower	Asian
U03	27	Female	Bachelor's degree	White
U04	37	Female	Bachelor's degree	Hispanic
U05	80	Female	Doctoral degree (Ph.D.)	White
U06	78	Female	Some college, no degree	White
U07	68	Female	Bachelor's degree	N/A
U08	50	Man	Bachelor's degree	White
U09	24	Female	Associate's degree	N/A
U10	46	Female	Master's degree	White
U11	72	Female	Bachelor's degree	Caucasian

Table 3. Formative study participant Demographics. Among all participants, only U10 reported reported having previous experiences with text-to-image AI systems; other participants reported no experience with text-to-image AI.

Major / Area of Study	Race	Gender	Current Program
Computer Science (13)	Asian (31)	Female (30)	Undergraduate (22)
Engineering (11)	White (9)	Male (13)	Master (18)
Design (10)	African American (3)	Non-binary (2)	Ph.D. (5)
Social Science (4)	Prefer not to answer (2)		
Public Policy (2)			
Philosophy (2)			
Math (2)			
English (1)			

Table 4. Overall demographics and background of our 45 user auditors from a U.S. university. Next to each demographic information, we include the number of the participants within that demographic group in parenthesis. Following our IRB protocol, we only collect minimum demographic data.

	Job title	Work experience relevant to our study	Years of experience	In Formative Study	In Evaluation Interview
P01	UX Researcher	Building a web-based application to engage users in flagging the potential biased and harmful behavior in a conversational agent	3 – 10	Yes	Yes
P02	ML Engineer	Engaging end users in rating the risk of representation harms of LLM applications	3 – 10	Yes	No
P03	UX Researcher	Building an application to engage users in auditing potential problematic behaviors of LLM models that have not yet been incorporated into products	3 – 10	Yes	No
P04	UX Researcher	Organizing focus groups to engage end users in testing their AI products and services	0 – 3	Yes	No
P05	Product Lead	Building an internal crowdsourcing tool for GenAI auditing	10+	Yes	Yes
P06	Technical Lead	Engaging users in auditing a range of AI products built by their customers	10+	Yes	No
P07	Data Scientist	Incorporating end users' feedback in measuring the representational harms of LLM applications	3 – 10	Yes	Yes
P08	Research Scientist	Human-centered evaluation for multimodal generative AI systems	0 – 3	No	Yes
P09	Research Engineer	Building an internal tool for engaging external domain experts in red-teaming generative AI	0 – 3	No	Yes
P10	Software Engineer	Building an internal evaluation pipeline for text-to-image generative AI	3 – 10	No	Yes
P11	Research Scientist	Training data curation for pre-trained large language and vision models	0 – 3	No	Yes

	Job title	Work experience relevant to our study	Years of experience	In Formative Study	In Evaluation Interview
P12	Research Scientist	Human-centered evaluation for multimodal generative AI systems	10+	No	Yes
P13	Data Scientist	Red teaming large language and vision models	3 – 10	No	Yes
P14	Software Engineer	Fairness testing for large language and vision models	0 – 3	No	Yes

Table 5. Summary of industry practitioner participants' backgrounds and relevant experience. All but one industry practitioner (P13, start up company) participants are from large technology companies.

Acknowledgements

Footnotes

1 https://beta.dreamstudio.ai/generate

2 Practitioners frame this activity differently depending on their organization, in some cases using terminology like "red-teaming".

3 Following prior work studying responsible AI practices in industry[121][122][123], to protect participants' confidentiality, we omit potentially identifying demographic details such as gender and age, and abstract our descriptions of participants' companies and roles.

4 We temporarily remove the link for double-blind review

5 See Section 5.1 for more details

6 https://www.prolific.com/

7 https://www.djangoproject.com/

8 https://replicate.com/

References

1. [a], [b], [c], [d]*Shen H, DeVos A, Eslami M, Holstein K (2021). "Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors". Proceedings of the ACM on Human-Computer Interaction. 5 (CSCW2): 1–29.*

2. [a], [b], [c], [d], [e], [f], [g], [h]*DeVos A, Dhabalia A, Shen H, Holstein K, Eslami M (2022). "Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior". In: Proceedings of the 2022 CHI conference on human factors in computing systems. pp. 1–19.*

3. [a], [b], [c], [d], [e]*Lam MS, Gordon ML, Metaxa D, Hancock JT, Landay JA, Bernstein MS (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, 34 pages. doi:10.1145/3555625.*

4. [a], [b], [c]*Mack KA, Qadri R, Denton R, Kane SK, Bennett CL (2024). "They only care to show us the wheelchair": disability representation in text-to-image AI models. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–23.*

5. [a], [b]*Shelby R, Rismani S, Rostamzadeh N (2024). "Generative AI in Creative Practice: ML-Artist Folk Theories of T2I Use, Harm, and Harm-Reduction". In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–17.*

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

7. [a], [b], [c], [d], [e], [f], [g], [h]*Deng WH, Guo B, Devrio A, Shen H, Eslami M, Holstein K (2023). "Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice". In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–18.*

8. [^]*Solyst J, Yang E, Xie S, Ogan A, Hammer J, Eslami M (2023). "The Potential of Diverse Youth as Stakeholders in Identifying and Mitigating Algorithmic Bias for a Future of Fairer AI". Proceedings of the ACM on Human-Computer Interaction. 7 (CSCW2): 1–27.*

9. [a], [b], [c]*Kieslich K, Diakopoulos N, Helberger N (2024). "Anticipating impacts: using large-scale scenario-writing to explore diverse implications of generative AI in the news environment". AI and Ethics. 2024:*

1–23.

10. a, b *Mun J, Jiang L, Liang J, Cheong I, DeCairo N, Choi Y, Kohno T, Sap M (2024). "Particip-ai: A democra tic surveying framework for anticipating future ai use cases, harms and benefits". In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 7: 997–1010.*

11. ^ *Maldaner MK, Deng WH, Hong J, Holstein K, Eslami M (2024). "MIRAGE: Multi-model Interface for Re viewing and Auditing Generative Text-to-Image AI". AAAI Conference on Human Computation and Cr owdsourcing (2024).*

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

13. a, b *Cabrera ÁA, Druck AJ, Hong JI, Perer A (2021). "Discovering and validating ai errors with crowdsourc ed failure reports". Proceedings of the ACM on Human-Computer Interaction. 5 (CSCW2): 1–22.*

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

15. a, b *Nushi B, Kamar E, Horvitz E (2018). "Towards accountable ai: Hybrid human-machine analyses for characterizing system failure". Proceedings of the AAAI Conference on Human Computation and Crowd sourcing. 6: 126–135.*

16. ^ *Ochigame R, Ye K (2021). "Search Atlas: Visualizing Divergent Search Results Across Geopolitical Borde rs". In: Designing Interactive Systems Conference 2021. pp. 1970–1983.*

17. ^ *Deng WH, Lam MS, Cabrera ÁA, Metaxa D, Eslami M, Holstein K (2023). "Supporting User Engagemen t in Testing, Auditing, and Contesting AI". In: Companion Publication of the 2023 Conference on Compu ter Supported Cooperative Work and Social Computing. 556–559.*

18. a, b *Ojewale V, Steed R, Vecchione B, Birhane A, Raji ID (2024). "Towards AI Accountability Infrastructur e: Gaps and Opportunities in AI Audit Tooling". arXiv preprint arXiv:2402.17861.*

19. a, b *Deng WH, Yurrita M, Díaz M, Suh J, Judd N, Groves L, Shen H, Eslami M, Holstein K (2024). "Respons ible Crowdsourcing for Responsible Generative AI: Engaging Crowds in AI Auditing and Evaluation". Pr oceedings of the AAAI Conference on Human Computation and Crowdsourcing. 12: 148--150.*

20. a, b, c *Birhane A, Steed R, Ojewale V, Vecchione B, Raji ID (2024). "AI auditing: The broken bus on the roa d to AI accountability". In: 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTM L). IEEE, 612–643.*

21. ^ *Bandy J (2021). "Problematic machine behavior: A systematic literature review of algorithm audits". P roceedings of the ACM on Human-Computer Interaction. 5 (CSCW1): 1–34.*

22. <u>^</u>*Weidinger L, Uesato J, Rauh M, Griffin C, Huang PS, Mellor J, Glaese A, Cheng M, Balle B, Kasirzadeh A, et al. (2022). "Taxonomy of risks posed by language models". In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 214–229.*

23. <u>^</u>*Weidinger L, Mellor J, Pegueroles BG, Marchal N, Kumar R, Lum K, Akbulut C, Diaz M, Bergman S, Rodriguez M, et al. (2024). "STAR: SocioTechnical Approach to Red Teaming Language Models". arXiv preprint arXiv:2406.11757.*

24. <u>^</u>*NIST (2023). "Artificial Intelligence Risk Management Framework". NIST. Available from: <u>https://www.nist.gov/itl/ai-risk-management-framework</u>.*

25. <u>a</u>, <u>b</u>*The White House (2023). "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence". The White House. Available from: <u>https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/</u>.*

26. <u>^</u>*Microsoft (2023). "Planning red teaming for large language models (LLMs) and their applications - Azure OpenAI Service". Available from: <u>https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming</u>.*

27. <u>^</u>*OpenAI (2023). GPT-4 System Card. Mar. Available from: <u>https://cdn.openai.com/papers/gpt-4-system-card.pdf</u>.*

28. <u>^</u>*OpenAI (2022). "OpenAI: Our approach to alignment research". <u>https://openai.com/blog/our-approach-to-alignment-research/</u>.*

29. <u>^</u>*Open AI (2022). "ChatGPT Feedback Contest: Official Rules". Open AI. Available from: <u>https://cdn.openai.com/chatgpt/ChatGPT_Feedback_Contest_Rules.pdf</u>.*

30. <u>^</u>*Anthropic (2023). "Model card and evaluations for Claude Models". <u>https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf</u>.*

31. <u>^</u>*Noble SU. Algorithms of oppression: How search engines reinforce racism. NYU Press; 2018.*

32. <u>^</u>*Asplund J, Eslami M, Sundaram H, Sandvig C, Karahalios K (2020). "Auditing race and gender discrimination in online housing markets". In: Proceedings of the International AAAI Conference on Web and Social Media. 14: 24–35.*

33. <u>a</u>, <u>b</u>*Sweeney L (2013). "Discrimination in online ad delivery". Queue. 11 (3): 10–29.*

34. <u>^</u>*Prates MOR, Avelar PH, Lamb LC (2020). "Assessing gender bias in machine translation: a case study with google translate". Neural Computing and Applications. 32 (10): 6363–6381.*

35. [a, b]Buolamwini J, Gebru T (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: Conference on fairness, accountability and transparency. PMLR, pp. 77–91.

36. [^]Hannak A, Soeller G, Lazer D, Mislove A, Wilson C (2014). "Measuring price discrimination and steering on e-commerce web sites". In: Proceedings of the 2014 Conference on Internet Measurement Conference. pp. 305–318.

37. [^]Holstein K, McLaren BM, Aleven V (2019). "Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms". In: International Conference on Artificial Intelligence in Education. Springer; 2019. p. 157–171.

38. [a, b]Sandvig C, Hamilton K, Karahalios K, Langbort C (2014). "Auditing algorithms: Research methods for detecting discrimination on internet platforms". Data and Discrimination: Converting Critical Concerns into Productive Inquiry. 22 (2014): 4349–4357.

39. [a, b, c, d]Metaxa D, Park JS, Robertson RE, Karahalios K, Wilson C, Hancock J, Sandvig C, et al. (2021). "Auditing algorithms: Understanding algorithmic systems from the outside in". Foundations and Trends® in Human--Computer Interaction. 14 (4): 272–344.

40. [^]Kaufmann N, Schulze T, Veit D (2011). "More than fun and money. worker motivation in crowdsourcing--a study on mechanical turk".

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

42. [^]Eslami M, Vaccaro K, Karahalios K, Hamilton K (2017). "Be careful; Things can be worse than they appear – Understanding biased algorithms and users' behavior around them in rating platforms". Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017. AAAI Press. pp. 62–71.

43. [^]Eslami M, Vaccaro K, Lee MK, Elazari Bar On A, Gilbert E, Karahalios K (2019). "User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms". Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019: 1–14. doi:10.1145/3290605.3300724.

44. [^]Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020). "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing". In: Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020:33-44.

45. [a, b, c]Chowdhury R, Williams J (2021). "Introducing Twitter's first algorithmic bias bounty challenge". URl: https://blog. twitter. com/engineering/en_us/topics/insights/2021/algorithmic-bias-bountychallen

ge. https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bountychalleng e.

46. ∧Zhang L, Liao X, Yang Z, Gao B, Wang C, Yang Q, Li D (2024). "Partiality and Misconception: Investigat ing Cultural Representativeness in Text-to-Image Models". In: Proceedings of the CHI Conference on H uman Factors in Computing Systems. pp. 1–25.

47. ∧Mim NJ, Nandi D, Khan SS, Dey A, Ahmed SI (2024). "In-Between Visuals and Visible: The Impacts of T ext-to-Image Generative AI Tools on Digital Image-making Practices in the Global South". In: Proceedi ngs of the CHI Conference on Human Factors in Computing Systems. pp. 1–18.

48. ∧Pistilli G (2022). "HuggingFace announcedthe new feature to flag any Model, Dataset, or Space on the Hub". https://twitter.com/GiadaPistilli/status/1571865167092396033?s=20&t=LRhhEu63s6ftPmtZdfz8C w.

49. ∧OpenAI (2022). "Open AI, ChatGPT". Available from: https://openai.com/blog/chatgpt/.

50. ∧Warkentin T, Woodward J (2022). "AI Test Kitchen". https://blog.google/technology/ai/join-us-in-th e-ai-test-kitchen/.

51. ∧DEF CON (2024). "Generative Red Team 2". DEF CON. Available from: https://grt.aivillage.org/.

52. ∧McGregor S, Ettinger A, Judd N, Albee P, Jiang L, Rao K, Smith W, Longpre S, Ghosh A, Fiorelli C, et al. (2024). "To Err is AI: A Case Study Informing LLM Flaw Reporting Practices". arXiv preprint arXiv:2410. 12104.

53. ∧Groves L, Peppin A, Strait A, Brennan J (2023). "Going public: the role of public participation approach es in commercial AI labs". In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 1162–1173.

54. ∧Feffer M, Sinha A, Deng WH, Lipton ZC, Heidari H (2024). "Red-teaming for generative ai: Silver bulle t or security theater?". In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 7: 421–43 7.

55. a, bVaughan JW (2017). "Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Le arning Research." J. Mach. Learn. Res.. 18 (1): 7026–7071.

56. a, bBigham JP, Bernstein MS, Adar E (2015). "Human-computer interaction and collective intelligence". Handbook of collective intelligence. 57. Mit Press Cambridge, MA, USA.

57. a, b, c, dKittur A, Nickerson JV, Bernstein M, Gerber E, Shaw A, Zimmerman J, Lease M, Horton J (2013). "The future of crowd work". In: Proceedings of the 2013 conference on Computer supported cooperative work. 1301–1318.

58. ^Sindlinger TS (2010). "Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business". *American Journal of Health-system Pharmacy*. **67**: 1565-1566.

59. a, bChilton LB, Little G, Edge D, Weld DS, Landay JA (2013). "Cascade: Crowdsourcing taxonomy creation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pp. 1999–2008.

60. a, b, c, dDow S, Kulkarni A, Klemmer S, Hartmann B (2012). "Shepherding the crowd yields better work". In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*. p. 1013–1022.

61. a, b, c, dChung JJY, Song JY, Kutty S, Hong S, Kim J, Lasecki WS (2019). "Efficient elicitation approaches to estimate collective crowd answers". *Proceedings of the ACM on Human-Computer Interaction*. **3** (CSCW): 1–25.

62. a, b, cMorris MR, Horvitz E (2007). "SearchTogether: an interface for collaborative web search". In *Proceedings of the 20th annual ACM symposium on User interface software and technology*. pp. 3–12.

63. a, bSiangliulue P, Arnold KC, Gajos KZ, Dow SP (2015). "Toward collaborative ideation at scale: Leveraging ideas from others to generate more creative and diverse ideas". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. pp. 937–945.

64. a, b, c, dKohn NW, Smith SM (2011). "Collaborative fixation: Effects of others' ideas on brainstorming". *Applied Cognitive Psychology*. **25** (3): 359–371.

65. ^Freyne J, Farzan R, Brusilovsky P, Smyth B, Coyle M (2007). "Collecting community wisdom: integrating social search & social navigation". In: *Proceedings of the 12th international conference on Intelligent user interfaces*. pp. 52–61.

66. a, bAmershi S, Morris MR (2008). "CoSearch: a system for co-located collaborative web search". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. pp. 1647–1656.

67. a, b, cWu T, Ribeiro MT, Heer J, Weld DS (2019). "Errudite: Scalable, reproducible, and testable error analysis". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pp. 747--763.

68. a, b, cBernstein MS, Little G, Miller RC, Hartmann B, Ackerman MS, Karger DR, Crowell D, Panovich K (2010). "Soylent: a word processor with a crowd inside". In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 2010. p. 313–322.

69. ^Bianchi F, Kalluri P, Durmus E, Ladhak F, Cheng M, Nozza D, Hashimoto T, Jurafsky D, Zou J, Caliskan A (2023). "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. p. 1493–1504.

70. <u>a</u>, <u>b</u>, <u>c</u>, <u>d</u>*Luccioni AS, Akiki C, Mitchell M, Jernite Y (2023). "Stable bias: Analyzing societal representations in diffusion models". arXiv preprint arXiv:2303.11408.*

71. <u>^</u>*Naik R, Nushi B (2023). "Social biases through the text-to-image generation lens". In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 786–808.*

72. <u>^</u>*Shelby R, Rismani S, Henne K, Moon A, Rostamzadeh N, Nicholas P, Yilla-Akbari N, Gallegos J, Smart A, Garcia E, et al. (2023). "Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction". In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 723–741.*

73. <u>^</u>*Campbell S, Greenwood M, Prior S, Shearer T, Walkem K, Young S, Bywaters D, Walker K (2020). "Purposive sampling: complex or simple? Research case examples". Journal of research in Nursing. 25 (8): 652–661.*

74. <u>^</u>*Gordon ML, Zhou K, Patel K, Hashimoto T, Bernstein MS (2021). "The disagreement deconvolution: Bringing machine learning performance metrics in line with reality". In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–14.*

75. <u>a</u>, <u>b</u>*Pirolli P, Card S (1999). "Information foraging." Psychological review. 106 (4): 643.*

76. <u>^</u>*Ling Lo M. Variation theory and the improvement of teaching and learning. Göteborg: Acta Universitatis Gothoburgensis; 2012.*

77. <u>a</u>, <u>b</u>, <u>c</u>*Gero KI, Swoopes C, Gu Z, Kummerfeld JK, Glassman EL (2024). "Supporting Sensemaking of Large Language Model Outputs at Scale". In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–21.*

78. <u>a</u>, <u>b</u>*Cosley D, Frankowski D, Terveen L, Riedl J (2007). "SuggestBot: using intelligent task routing to help people find work in wikipedia". In: Proceedings of the 12th international conference on Intelligent user interfaces. pp. 32–41.*

79. <u>^</u>*Luther K, Tolentino J-L, Wu W, Pavel A, Bailey BP, Agrawala M, Hartmann B, Dow SP (2015). "Structuring, aggregating, and evaluating crowdsourced design critique". In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing. pp. 473–485.*

80. <u>^</u>*Dourish P, Lawrence C, Leong TW, Wadley G (2020). "On being iterated: The affective demands of design participation". In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–11.*

81. <u>a</u>, <u>b</u>*Pierre J, Crooks R, Currie M, Paris B, Pasquetto I (2021). "Getting Ourselves Together: Data-centered participatory design research & epistemic burden". In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–11.*

82. a, b, c*Wang ZJ, Kulkarni C, Wilcox L, Terry M, Madaio M (2024). "Farsight: Fostering Responsible AI Aw areness During AI Application Prototyping". In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–40.*

83. ^*Deng WH, Nagireddy M, Lee MSA, Singh J, Wu ZS, Holstein K, Zhu H (2022). "Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits". In: 2022 ACM Conference on Fairness, Accounta bility, and Transparency. ACM; Seoul Republic of Korea. p. 473-484. doi:10.1145/3531146.3533113.*

84. ^*Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J (2020). "Interpreting interpreta bility: 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.*

85. ^*Braun V, Clarke V (2019). "Reflecting on reflexive thematic analysis". Qualitative research in sport, exe rcise and health. 11 (4): 589–597.*

86. ^*Chen QZ, Zhang AX (2023). "Judgment Sieve: Reducing uncertainty in group judgments through interv entions targeting ambiguity versus disagreement". Proceedings of the ACM on Human-Computer Inter action. 7 (CSCW2): 1–26.*

87. ^*Kuo TS, Halfaker AL, Cheng Z, Kim J, Wu MH, Wu T, Holstein K, Zhu H (2024). "Wikibench: Community -driven data curation for AI evaluation on Wikipedia". In Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–24.*

88. ^*Kingsley S, Zhi J, Deng WH, Lee J, Zhang S, Eslami M, Holstein K, Hong JI, Li T, Shen H (2024). "Investi gating What Factors Influence Users' Rating of Harmful Algorithmic Bias and Discrimination". In: Proce edings of the AAAI Conference on Human Computation and Crowdsourcing. 12: 75–85.*

89. ^*Passi S, Jackson SJ (2018). "Trust in data science: Collaboration, translation, and accountability in corp orate data science projects". Proceedings of the ACM on Human-Computer Interaction. 2 (CSCW): 1–28.*

90. a, b*Kraut RE, Resnick P (2011). "Encouraging contribution to online communities". Building successful o nline communities: Evidence-based social design. MIT Press Cambridge, MA: 21–76.*

91. ^*Burke M, Kraut R, Marlow C (2011). "Social capital on Facebook: Differentiating uses and users". In Pro ceedings of the SIGCHI conference on human factors in computing systems. pp. 571–580.*

92. ^*Lampe C, Wash R, Velasquez A, Ozkaya E (2010). "Motivations to participate in online communities". I n: Proceedings of the SIGCHI conference on Human factors in computing systems. 1927-1936.*

93. ^*Hara K, Adams A, Milland K, Savage S, Callison-Burch C, Bigham JP (2018). "A data-driven analysis o f workers' earnings on Amazon Mechanical Turk". Proceedings of the 2018 CHI conference on human fa ctors in computing systems. 1–14.*

94. ^Dosono B, Semaan B (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. pp. 1–13.

95. ^Gray ML, Suri S. Ghost work: How to stop Silicon Valley from building a new global underclass. Eamon Dolan Books; 2019.

96. ^Steiger M, Bharucha TJ, Venkatagiri S, Riedl MJ, Lease M (2021). "The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support". In: Proceedings of the 2021 CHI conference on human factors in computing systems. pp. 1–14.

97. ^Reynante B, Dow SP, Mahyar N (2021). "A framework for open civic design: Integrating public participation, crowdsourcing, and design thinking". Digital Government: Research and Practice. 2 (4): 1–22.

98. ^Arnstein SR (1969). "A ladder of citizen participation". Journal of the American Institute of planners. 35 (4): 216–224.

99. ^Wang C, Deng WH, Hong J, Holstein K, Eslami M (2024). "Designing a Crowdsourcing Pipeline to Verify Reports from User AI Audits". AAAI Conference on Human Computation and Crowdsourcing.

100. a, b, c Bu\u{c}inca Z, Pham CM, Jakesch M, Ribeiro MT, Olteanu A, Amershi S (2023). "AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms". arXiv preprint arXiv:2306.03280.

101. a, b, c Rastogi C, Ribeiro MT, King N, Nori H, Amershi S (2023). "Supporting human-ai collaboration in auditing llms with llms". In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. p. 913-926.

102. ^Pang RY, Reinecke K (2023). "Anticipating Unintended Consequences of Technology Using Insights from Creativity Support Tools". arXiv preprint arXiv:2304.05687.

103. ^Park JS, Popowski L, Cai C, Morris MR, Liang P, Bernstein MS (2022). "Social simulacra: Creating populated prototypes for social computing systems". In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. pp. 1–18.

104. ^Pang RY, Santy S, Just R, Reinecke K (2024). "BLIP: Facilitating the Exploration of Undesirable Consequences of Digital Technologies". In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–18.

105. ^Zaydi M, Maleh Y (2024). "EMPOWERING RED TEAMS WITH GENERATIVE AI: TRANSFORMING PENETRATION TESTING THROUGH ADAPTIVE INTELLIGENCE". EDPACS. 2024: 1–26.

106. ^Dev S, Jha A, Goyal J, Tewari D, Dave S, Prabhakaran V (2023). "Building stereotype repositories with llms and community engagement for scale and depth". Cross-Cultural Considerations in NLP@ EACL. 84

(2023).

107. [a, b, c]*Chiu YY, Jiang L, Antoniak M, Park CY, Li SS, Bhatia M, Ravi S, Tsvetkov Y, Shwartz V, Choi Y (2024). "CulturalTeaming: AI-Assisted Interactive Red-Teaming for Challenging LLMs'(Lack of) Multicultural Knowledge". arXiv preprint arXiv:2404.06664.*

108. [a, b, c]*Radharapu B, Robinson K, Aroyo L, Lahoti P (2023). "Aart: Ai-assisted red-teaming with diverse data generation for new llm-powered applications". arXiv preprint arXiv:2311.08592.*

109. [^]*Cinelli M, De Francisci Morales G, Galeazzi A, Quattrociocchi W, Starnini M (2021). "The echo chamber effect on social media". Proceedings of the National Academy of Sciences. 118 (9): e2023301118.*

110. [^]*Liu MX, Wu T, Chen T, Li FM, Kittur A, Myers BA (2024). "Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models". In Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–26.*

111. [^]*Jiang P, Rayan J, Dow SP, Xia H (2023). "Graphologue: Exploring large language model responses with interactive diagrams". In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. p. 1–20.*

112. [^]*Chen NC, Suh J, Verwey J, Ramos G, Drucker S, Simard P (2018). "AnchorViz: Facilitating classifier error discovery through interactive semantic data exploration". In: Proceedings of the 23rd International Conference on Intelligent User Interfaces. 269–280.*

113. [^]*Madaio M, Egede L, Subramonyam H, Vaughan JW, Wallach H (2021). "Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support". arXiv preprint arXiv:2112.05675 (2021).*

114. [^]*DeVrio A, Eslami M, Holstein K (2024). "Building, Shifting, & Employing Power: A Taxonomy of Responses From Below to Algorithmic Harm". In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 1093–1106.*

115. [^]*Young M, Magassa L, Friedman B (2019). "Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents". Ethics and Information Technology. 21 (2): 89–103.*

116. [^]*Miceli M, Posada J (2021). "Wisdom for the crowd: discoursive power in annotation instructions for computer vision". arXiv preprint arXiv:2105.10990.*

117. [^]*Irani LC, Silberman MS (2013). "Turkopticon: Interrupting worker invisibility in amazon mechanical turk". In Proceedings of the SIGCHI conference on human factors in computing systems. pp. 611–620.*

118. ^Salehi N, Irani LC, Bernstein MS, Alkhatib A, Ogbe E, Milland K (2015). "We are dynamo: Overcoming s talling and friction in collective action for crowd workers". In: *Proceedings of the 33rd annual ACM conf erence on human factors in computing systems.* 2015: 1621–1630.

119. ^Vincent N, Li H, Tilly N, Chancellor S, Hecht B (2021). "Data leverage: A framework for empowering th e public in its relationship with technology companies". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* pp. 215--227.

120. ^Longpre S, Kapoor S, Klyman K, Ramaswami A, Bommasani R, Blili-Hamelin B, Huang Y, Skowron A, Yong ZX, Kotha S, et al. (2024). "A safe harbor for ai evaluation and red teaming". *arXiv preprint arXiv: 2403.04893.*

121. ^Madaio MA, Stark L, Wortman Vaughan J, Wallach H (2020). "Co-designing checklists to understand o rganizational challenges and opportunities around fairness in ai". In: *Proceedings of the 2020 CHI Conf erence on Human Factors in Computing Systems.* pp. 1–14.

122. ^Rakova B, Yang J, Cramer H, Chowdhury R (2021). "Where responsible AI meets reality: Practitioner pe rspectives on enablers for shifting organizational practices". *Proceedings of the ACM on Human-Compu ter Interaction.* **5** (CSCW1): 1–23.

123. ^Deng WH, Yildirim N, Chang M, Eslami M, Holstein K, Madaio M (2023). "Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice". In: *Proceedin gs of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* pp. 705–716.

Declarations