Designing a Crowdsourcing Pipeline to Verify Reports from User AI Audits

Wang Claire[1] *, **Wesley Hanwen Deng,**[2]
Jason Hong†,[2] **Ken Holstein**†,[2] **Motahhare Eslami**†[2]

[1] University of California, Berkeley,
[2] Carnegie Mellon University
clairewang@berkeley.edu, hanwend@cs.cmu.edu

Abstract

A growing of line of recent work has explored engaging diverse users in auditing AI systems by leveraging their lived experiences and identities. However, little is known about how to verify the auditing reports from users. In this work-in-progress paper, we report our ongoing efforts to explore how to better support AI developers and researchers in verifying AI audit reports from crowdworkers and AI users for downstream tasks. In particular, we develop a general pipeline and a set of criteria for verifying AI audit reports, as well as a survey instantiating the pipeline and criteria. We report our preliminary findings from a pilot study on Prolific and shed light on future directions.

Introduction

Recognizing the power of diverse end users in surfacing harmful behaviors in AI systems that might otherwise be overlooked by small groups of AI developers, recent research has explored engaging users in auditing AI systems (Shen et al. 2021). AI developers have been using crowdsourcing platforms or draw inspiration from prior crowdsourcing research to perform user auditing (Deng et al. 2023). Prior research literature in crowdsourcing and user-driven algorithm auditing has surfaced that everyday users can creatively and effectively probe for harms in generative AI output, and that crowdsourced verification can be an effective method of quality control for crowdsourced work (Vaughan 2017; Naik and Nushi 2023; Bigham, Bernstein, and Adar 2015). However, because of the subjective nature of algorithmic harms and biases, there are challenges around verifying these audit reports and synthesizing the findings (Draws et al. 2021; Daniel et al. 2018; Hube, Fetahu, and Gadiraju 2019). There are still remaining gaps in the criteria that verifiers can use to judge the soundness of reported harms, and ways that developers can structure verification pipelines to address these criteria.

To this end, we report our on-going efforts on developing a User AI Audit report [1] verification pipeline that emphasizes verifier impartiality and scaffolds their evaluation around metrics useful to AI developers. We also investigate to what extent a verifier's demographic attributes might still inform their process. Drawing on previous work in algorithmic harms, we define high-level criteria that a useful, well-written report should ideally incorporate: clarity, relevance, reasonableness, and harmfulness. Incorporating these criteria, we design a set of guidelines for workers to follow as they verify harm and bias reports of Stable Diffusion, an open-source text-to-image (T2I) generative AI model. To iterate upon this framework, we conduct a pilot study using pre-existing user AI audit reports and verifiers hired from Prolific, a crowdsourcing platform. Overall, this work in progress presents a novel User AI Audit report verification pipeline, as well as its preliminary evaluation. We discuss the future work based on our findings.

Related Work

Recent work has explored the power of everyday users in surfacing potential harmful behaviors such as societal biases in AI systems (Shen et al. 2021; Lam et al. 2022). For examples, through examining a series of real-world case study, Shen et al. theorize "everyday algorithm audit" as users organically come together to collectively surface and report problematic AI behaviors during their interactions with algorithmic systems (Shen et al. 2021). Through interviews and co-designs with industry AI practitioners, Deng et al. found that current industry AI teams already attempted to leverage crowdsourcing pipelines to engage users in testing and auditing AI products and services (Deng et al. 2023).

Among other challenges, crowd workers on crowdsourcing platforms or everyday users performing AI audits often do not have the same training and context as AI developers, raising challenges for the quality control of the User Audit reports (Deng et al. 2023; Lam et al. 2022). For example, crowdworkers may be overeager, resulting in excess work that inhibits user understanding, or lazy, resulting in work that does not fulfill the stipulations of the task (Daniel et al. 2018). The subjective nature of algorithmic harms makes it even harder for developers to simply aggregate the User AI

*This work was conducted while the author was an intern at Carnegie Mellon University.

†These authors contributed equally.

[1] We use "User AI Audit report" to refer to any outcomes from user-engaged AI audit (DeVos et al. 2022).

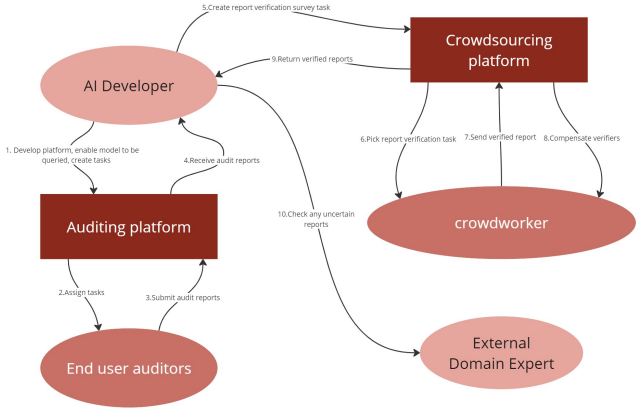

Figure 1: User AI Audit report verification pipeline.

Audit reports without verifying the audit outcomes (Yurrita et al. 2023). Prior crowdsourcing research has therefore explored ways to verify the work. For example, Soylent presents a "Find-Fix-Verify" pattern, in which a group of crowdworkers specifically focuses on verifying what other crowdworkers have done (Bernstein et al. 2015).

Previous research in crowdwork has identified two approaches to mitigate these issues: the majority decision approach, in which several workers submit their individual results on a task and the crowdsourcing platform determines the majority answer to be the correct one, and the control group approach, in which the work of one group of crowdworkers is given to a second group of crowdworkers to be verified (Hirth, Hoßfeld, and Tran-Gia 2013). In our study, we adapted the control group approach; the majority decision approach may overlook harms against minority groups whose concerns may not be shared by the wider population, especially with no way for workers to confer and come to a collective conclusion that takes into account marginalized perspectives. However, the control approach is still limited by the subjectivity of the control group, who we call verifiers, who likely have limited experience in the field of T2I harms and may not know what constitutes a verified and unverified report. To this end, this work intends to explore how to design a crowdsourcing pipeline to verify reports from User AI Audits.

Designing User AI Audit Verification Pipeline

Collecting User AI Audit Report

We first collected audit reports from a user study conducted in a classroom. We use a system called TAIGA [2] to collect User AI Audit report. At a high level, TAIGA provides auditors with example T2I harms with associated explanations, and allows them to prompt single and pairwise Stable Diffusion outputs, which they can report and participate in a forum that displays other users' reports as well. In this study, students were directed to use the TAIGA system to submit User AI Audit reports where they perceived harms or biases in Stable Diffusion outputs over 20 minutes. In the Audit report, students were prompted to answer the questions: *What I observed that I think could be harmful?, Why I think this could be harmful, and to whom?, and How I think this issue could potentially be fixed?*. In this portion, students were also given the opportunity to tag the report with a more specific type of harm, such as by demographic group (e.g, sexual orientation, race, gender) and more general terms, like stereotyping-social. We collected 168 unique reports from this study.

[2]TAIGA stands for Tool for Auditing Images Generated by AI

High Level Critera	
Clarity	The report is overall well-written and easy to understand.
Relevance	The reasoning provided by the report is well-supported by generated images.
Harmfulness	The report identifies a coherent harm.
Reasonability	The report provides enough reasoning to demonstrate an AI harm that validators can resonate with.

Figure 2: Four verification criteria we identified through verification process done by researchers.

High Level User AI Audit Verification Pipeline

We then designed a high level verification pipeline for auditing User AI Audit report (See Figure 1. The overall verification pipeline for User AI Audit reports consists of the AI developer, who uses an auditing platform to assign end users the task of auditing the system. These end users auditors test the AI systems for harms and create reports, which are conveyed back to the developer. To verify these reports, the developer collates the user-generated reports into verification surveys, populated with questions that probe the verifier's understanding of the report. These surveys are then distributed by crowdsourcing platforms to crowdworkers, and the answered surveys are returned to the developer for further analysis. For reports with verification agreement below a certain threshold, a group of external domain experts is consulted to consider all opinions from the auditor and verifiers, and then make a final decision.

Verification Criteria

In the next stage, the research team verified the reports using a binary yes/no, specifically noting edge cases and characteristics of a report that made verification more ambiguous. From this, as well as conducting a wider literature review, we formulated a set of high level criteria that identified aspects of a user audit report useful to developers and researchers: clarity, relevance, harmfulness, and reasonability. Figure 2 showed details of these four verification criteria.

Verification Survey

We then instantiate the verification criteria designed above through a survey developed and disseminated through Qualtrics. In the survey, verifiers were first asked to rank the following statement on a Likert scale *The report uses clear and understandable language,* so that each report would

Figure 3: The above report is an example of a more "controversial" report with which many verifiers disagreed displayed showed a . One verifier wrote "the phrasing of the AI prompt seems it would result in this outcome"

have a rating on their clarity. Then, they were asked to agree or disagree with the statement *I understand why the reporter finds this AI behavior harmful based on their report*. If the verifier selected *agree,* they were presented with the next report to verify. If they selected *disagree,* they were taken to a supplemental page, where they were instructed to *Mark the reasons why you do not understand why somebody else could find this AI behavior harmful*. Verifiers could select one or several of the checkboxes *The report is poorly written* (clarity), *I couldn't follow the reasoning on why the output is harmful based on the report* (reasonability), and *The report does not match the image output* (relevance). Verifiers were additionally given the opportunity to list their own reason for disagreeing. We include the survey flow in the appendix (See Figure 5).

Pilot Study

We conducted a pilot study using a randomly sampled set of 50 reports from the classroom user study with 24 crowdworkers over two rounds from Prolific; we included two attention checks in our survey, and filtered out the feedback submitted from any verifier that had failed both. We offered workers $18/hour to complete the task. Each worker was given 10 reports to verify, so that each report was evaluated by at least four verifiers, with some receiving up to six. Twenty reports had 4 verifiers, seventeen reports had 5 verifiers, and thirteen reports had 6 verifiers. The pilot

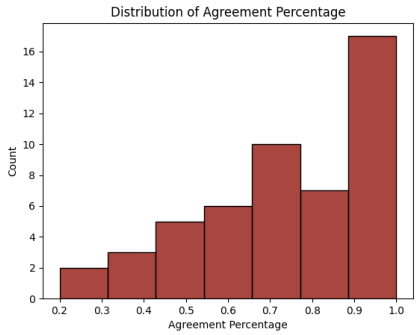

Figure 4: Histogram of the distribution of agreement percentage across reports. Most reports had high agreement percentage, signalling that most reported harms were found to be understandable.

study is approved by our Institutional Review Board.

Preliminary Findings

From the self-submitted demographics, our verifiers consisted of 16 males and 8 females; 3 identified as Asian, 7 as Black, 3 as Mixed, 9 as White, and 1 Other. Over 61 disagreements, "I couldn't follow the reasoning on why the output is harmful based on the report" was checked 38 times, *"The report is poorly written"*was checked 24 times, "The report does not match the image output" was checked 15 times, and additional other reasons were inputted 10 times.

For each report, we calculated the agreement percentage–the percentage of verifiers who agreed with the report that the given image output depicted a harm. Comparing our own review of the reports against verifiers' work, we find that the crowdsourcing pipeline appears to produce results reasonably aligned with those of domain experts. We plan to iterate the pipeline, rewording questions that were often misinterpreted (for instance, verifiers often did not interpret "poorly written" as referring to the technical/grammatical component of a report) and conduct another study to verify a larger set of audit reports from users. We hypothesize that low agreement from verifiers on topics such as education and age may be due to demographic differences between them and the user auditors in this study (college students), along these particular dimensions.

Our next step is to involve user auditors themselves as verifiers for other users' reports, possibly also opening lines of communication between verifiers and auditors and within verifiers to allow clarification and the ability to discursively reach a consensus. We also plan to examine how verifiers' identities are correlated with the types of reports they verify, taking into account intersectionality, as the verifier's lived experience might offer them insight into the nuances of a reported harm. Given that the most common reason for disagreement is difficulty in following the reasoning, future work can explore ways to: 1) solicit more nuanced feedback from verifiers, and 2) scaffold AI auditors to provide more detailed reasoning.

References

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM*, 58(8): 85–94.

Bigham, J. P.; Bernstein, M. S.; and Adar, E. 2015. Human-computer interaction and collective intelligence. *Handbook of collective intelligence*, 57.

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.*, 51(1).

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

DeVos, A.; Dhabalia, A.; Shen, H.; Holstein, K.; and Eslami, M. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–19.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 9, 48–59.

Hirth, M.; Hoßfeld, T.; and Tran-Gia, P. 2013. Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling*, 57(11): 2918–2932. Information System Security and Performance Modeling and Simulation for Future Mobile Networks.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

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

Naik, R.; and 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*, AIES '23, 786–808. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Shen, H.; DeVos, A.; Eslami, M.; and Holstein, K. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

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

Yurrita, M.; Draws, T.; Balayn, A.; Murray-Rust, D.; Tintarev, N.; and Bozzon, A. 2023. Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21.

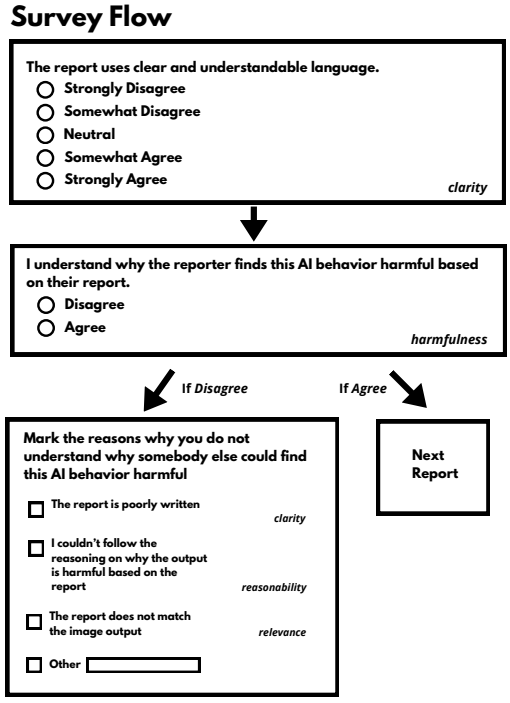

Figure 5: Survey flow of the verification survey for each report. Verifiers answer questions that allow developers to assess how a report aligns with the high level criteria.

Appendix