Organizing Crowd Audits to Detect Bias in Machine Learning

Wesley Hanwen Deng, Hong Shen, Adam Perer, Motahhare Eslami, Kenneth Holstein, Jason I. Hong

Carnegie Mellon University

Human-Computer Interaction Institute

How might we help industry practitioners combat ML bias?

Developing useful and usable ML fairness toolkits that attend to industry ML practitioners' on-the-ground need.

ML Fairness Toolkit developers' assumptions

Misalignment

ML practitioners' actual usage and desire

Practitioners are **already knowledgeable** in fair ML before using the toolkits.

ML Fairness toolkits are mainly used by **technical roles**.

Practitioners **misuse** toolkits based on **misconceptions** about ML fairness, such as *fairness through unawareness*.

ML Fairness work requires close **collaboration** among **cross-functional** roles (*technical, user-facing, business, compliance*)

Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. Wesley Deng, et al.. FAccT 2022;
Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. Wesley Deng, et al. FAccT 2023;
Zeno: An Interactive Framework for Behavioral Evaluation of Machine Learning. Ángel Alexander Cabrera, et al. CHI 2023.

Building effective tools to better support industry AI practitioners in better engage diverse end-users in auditing their AI products and services ("crowd audits")

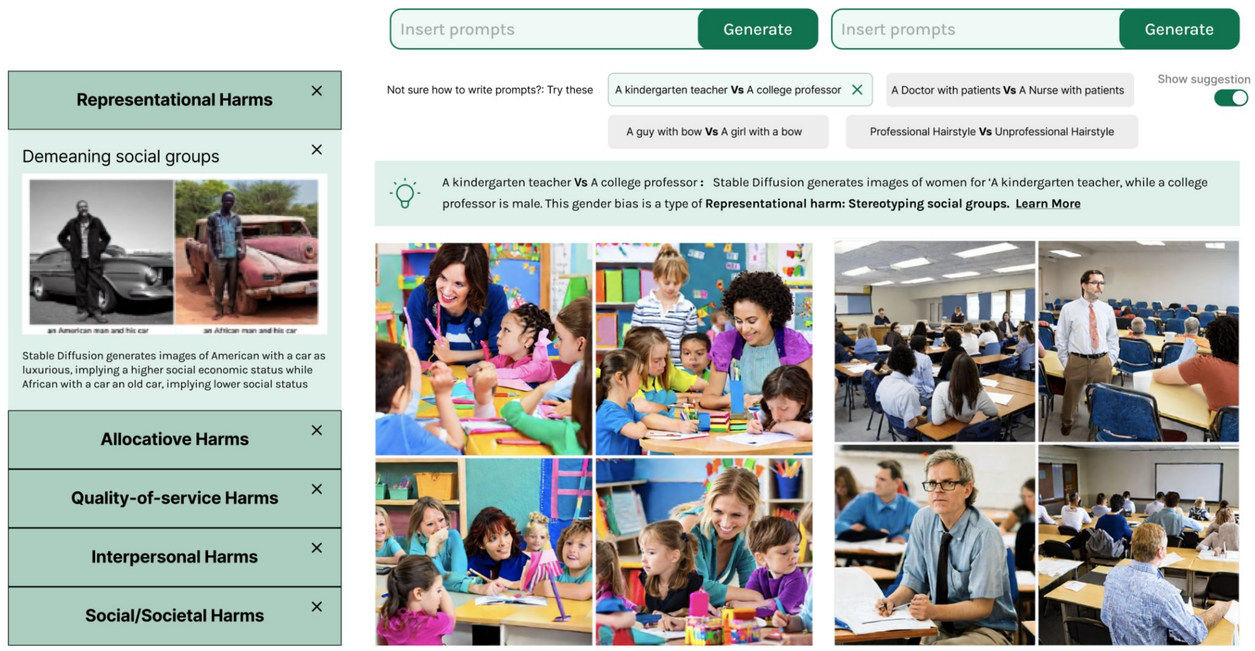
Identifying and recruiting the right group of auditors	**Incentivizing** sustained, high-quality user contribution	Effectively **scaffolding** users in auditing algorithms	**Deriving actionable insights** from user-engaged auditing

Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors Hong Shen, et al. CSCW 2021;
Discovering and Validating AI Errors With Crowdsourced Failure Reports. Ángel Alexander Cabrera, et al., CSCW 2021;
Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. Alicia DeVos, et al., CHI 2022;
Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. Wesley Deng, et al., CHI 2023.

How might we empower end-users in auditing generative AI?

WeAudit: An interactive platform to support "crowd audits"

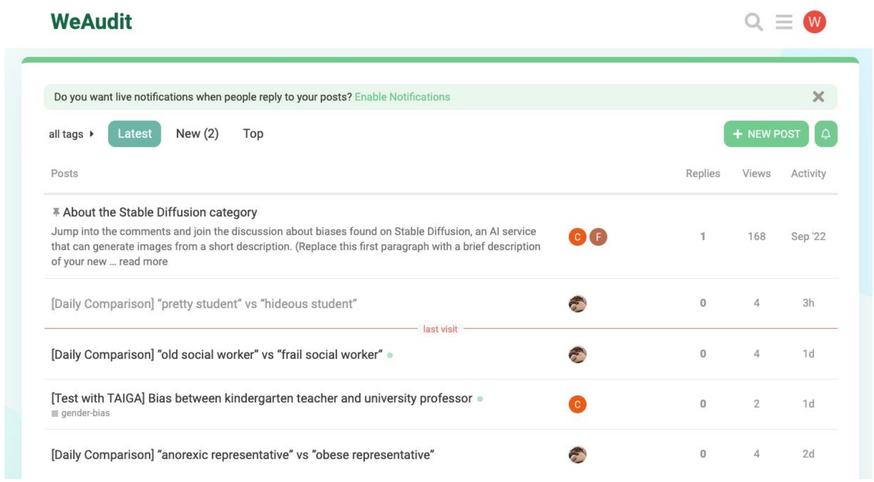
Step 1. *User Auditors* ***enter prompt pairs*** *to conduct audit.*



Socio-technical Harms Taxonomy *for text-to-image generative AI available on sidebar for reference. User auditors will also be directed to this taxonomy when exploring the suggestions, to enhance their understanding of potential harms.*

Step 2. *User Auditors* ***compare AI-generated images*** *side-by-side.*

Step 3. *User Auditors* ***submit an "audit report"*** *to the* ***"WeAudit Forum,"*** *where they can access other users' audit reports and engage in discussions and deliberations.*



Keep connected:

Wesley Deng hanwend@cs.cmu.edu @wes_deng

Jason Hong jasonh@cs.cmu.edu @jas0nh0ng

Try WeAudit:

https://forum.weaudit.org/