

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

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

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

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, highquality user contribution

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.

Organizing Crowd Audits to Detect Bias in Machine Learning

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







ML practitioners' actual usage and desire

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)

Effectively scaffolding users in auditing algorithms

Deriving actionable insights from userengaged auditing









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.







WeAudit: An interactive platform to support "crowd audits"



Insert prompts

A kindergarten teac



A kindergarten teacher Vs A college professor: Stable Diffusion generates images of women for 'A kindergarten teacher, while a college professor is male. This gender bias is a type of Representational harm: Stereotyping social groups. Learn More



Step 2. User Auditors compare Al-generated images side-by-side.

engage in discussions and deliberations.

WeAudit

Do you want live notifications when people reply to your posts? Enable Notifications		
Il tags 🕨 Latest New (2) Top		
Posts	R	eplies
About the Stable Diffusion category		
Jump into the comments and join the discussion about biases found on Stable Diffusion, an AI service that can generate images from a short description. (Replace this first paragraph with a brief description of your new read more	CF	1
Daily Comparison] "pretty student" vs "hideous student"	0	0
Daily Comparison] "old social worker" vs "frail social worker" 💿	0	0
Test with TAIGA] Bias between kindergarten teacher and university professor <pre>o</pre>	0	0
Daily Comparison] "anorexic representative" vs "obese representative"	0	0



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

Generate Insert prompts		Generate	
er Vs A college professor X	A Doctor with patients Vs A Nurse with patients	Show suggestic	
A girl with a bow	rofessional Hairstyle Vs Unprofessional Hairstyle		

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

	Q	W
		×
6	+ NEW P	ost 🎝
	Views	Activity
	168	Sep '22
	4	Зh
	4	1d
	2	1d
	4	2d

Keep connected:

Wesley	hanwend@cs.cmu.edu
Deng	@wes_deng
Jason	jasonh@cs.cmu.edu

@jas0nh0ng Hong

Try WeAudit:



https://forum.weaudit.org/