Uncovering Gender Biases in Human-AI Platforms

Siddharth D Jaiswal

Indian Institute of Technology Kharagpur, India
siddsjaiswal@kgpian.iitkgp.ac.in

Research Direction

Human-AI platforms are AI-based systems that humans directly interface with, for addressing some personal or social need. Examples range from media recommendation platforms (Netflix) and e-commerce websites (Amazon) to face (AWS Rekognition) and speech (OpenAI Whisper) recognition platforms. These platforms have now become ubiquitous. While the benefits of using such systems are well known, a wide variety of biases have also been reported against different stakeholders. For example- there have been reports of discrimination against dark-skinned people by Face Recognition Systems (FRSs) (Jaiswal et al. 2022), biases against non-binary shoppers (Jaiswal and Mukherjee 2022) and users of social media platforms (Jaiswal, Verma, and Mukherjee 2023). These biases have far-reaching consequences on society. It is essential to identify and mitigate these biases to ensure fair treatment for all individuals involved. To identify these biases, researchers perform (third-party) audit studies that probe the platforms with various inputs (standard & adversarial) and analyze the outputs along different demographic dimensions. To mitigate the biases, interventions are introduced in one or more of the following stages of model development– pre-processing (changes to the training dataset), in-processing (changes to the objective function) or post-processing (regularization/smoothing) steps. Such studies form the bedrock of Responsible AI.

Problem Statement: As part of my PhD studies, I am focusing on the broad problem of *gender related biases in Human-AI systems*, making a three-fold contribution to every part of the Responsible AI pipeline– (i) Adversarial Audits in different human-AI platforms to identify biases against binary & non-binary gender groups, (ii) Novel gender-inclusive datasets and (iii) Low-resource bias mitigation algorithms. The Human-AI systems that I am studying as part of my PhD thesis are– (a) Face Recognition Systems (FRSs), (b) e-commerce platforms, (c) text-based gender analyzers and (d) Vision Language Models (VLMs). My thesis is divided into three parts– (i) Adversarial audit of FRSs (Jaiswal and Mukherjee 2022; Jaiswal, Verma, and Mukherjee 2024; Jaiswal et al. 2024) for binary genders, (ii) Data-centric (Jaiswal et al. 2024) and model-centric (current work) bias mitigation in FRSs for binary genders, and (iii) Audit of non-binary gender bias in other Human-AI systems– visual-search enabled e-commerce (Jaiswal and Mukherjee 2022), text-based gender analyzers (Jaiswal, Verma, and Mukherjee 2023) and VLMs (current work).

Research Questions and Contributions

Since the beginning of my PhD, I have addressed the following research questions–

RQ1. Adversarial audits for FRSs: Audits probe systems with various types of inputs to identify discrimination or biases. Many studies have performed audits that have exposed biases against marginalized groups like Black females. These audit strategies use standard benchmark inputs composed of face images which are then used to evaluate model performance and discrimination for tasks like gender, age and emotion prediction.

Gaps identified: (1) Real world inputs are not as simple as benchmark datasets. Most cameras deployed in the wild are exposed to natural elements like rain, dust, etc and in a post pandemic world, most people wear masks in public, (2) Existing audit studies don't evaluate for model robustness under realistic conditions.

Our Contributions: We have developed a new audit strategy, called an "adversarial" audit. The adversarial inputs simulate realistic conditions like blurring of the camera lens, face masks, noisy images from social media, etc. (addresses Gap (1)). Next, to address Gap (2), these adversarial inputs also allow us to study the impact on model robustness from the lens of discrimination. While it is expected that the model accuracy will be impacted under adversarial scenarios, it is more important to know the impact on the different social or intersectional groups [1]. In Jaiswal et al. (2022), we audit three commercial FRSs and identify that biases for gender, age and smile classification against dark-skinned females increase manifold under realistic adversarial conditions, on 5 benchmark datasets. In Jaiswal, Verma, and Mukherjee (2024), we audit thirteen FRSs (commercial and open-source) for face verification and identification with masked faces and discover similar biases against dark-skinned females. A survey with human participants shows

[1] Any group formed by combining race and gender labels– white male, black female, etc.

they are similarly biased.

RQ2. Bias mitigation solutions: Bias mitigation can be performed at one of the three intervention stages– pre-processing, in-processing or post-processing. In my PhD, I am focusing on the pre- and in-processing stages using data-centric and model-centric solutions, respectively. All AI-based platforms need well-sampled and representative data to function in an unbiased manner. Thus, access to diverse, fair datasets is of utmost importance not just for model development but also audits. Moreover, even with diverse data, models may exhibit large discrimination and need redesign at an architectural level.

Gaps identified: (1) Existing datasets are highly imbalanced in terms of racial/ethnic and gender representations and the models trained on these datasets carry forward the biases, (2) A majority of existing datasets violate the privacy of regular citizens as their photos are collected from their social media profiles without their permission, (3) There is a lack of sufficient datasets from the Global South countries. As these countries are the hotbed of deployment for AI software, it is important to ensure that their citizens are fairly represented in the data, and, (4) There are very few adversarial datasets for training and auditing FRS models, (5) Existing model-centric bias mitigation solutions may be dataset-dependent and require large-scale changes to the model.

Our Contributions: We have developed a benchmark dataset (Jaiswal et al. 2024) that has representation from various ethnicities and races (addresses Gap (1)) and is composed of publicly famous individuals (addresses Gap (2)). Our dataset has more than 50% individuals belonging to the Global South countries (addresses Gap (3)) and has five adversarial variants (addresses Gap (4)). This large scale dataset (more than 40k images including the adversarial variants) will allow us to train, test and audit both commercial and open-source FRSs under standard and adversarial conditions. In my current work I am developing model-centric bias mitigation algorithms for FRSs by employing simple architectural changes– changing the number of layers, loss functions and using residual connections (addresses Gap (5)).

RQ3. Biases against non-binary individuals: Human-AI platforms like visual search enabled e-commerce, text-based gender analyzers and VLMs are used for various personal and professional purposes, directly interfacing with humans. Developers must design them with due acknowledgement to the minority groups who may use these platforms. Non-binary individuals are one such group who often face discrimination in society and if this discrimination is baked into the AI models deployed at scale around the world, these individuals may face unprecedented biases. Thus it is important to audit and fix existing platforms for such biases using sufficiently diverse datasets.

Gaps identified: We have identified the following major gaps in existing literature– (1) Existing audit studies rarely look at biases against non-binary individuals, even in domains where such datasets may be available and, (2) Most existing models are not designed with gender fluidity in mind, either due to developer oversight or deliberate choice.

Our Contributions: I have curated two novel datasets where more than 50% of the data points belong to the non-binary gender group– an image dataset of fashion wear for male, female and non-binary clothing (Jaiswal and Mukherjee 2022) and a text-dataset of Reddit and Tumblr comments/posts from self-declared non-binary individuals (Jaiswal, Verma, and Mukherjee 2023) that can be used to audit visual search enabled e-commerce and text-based gender analyzers (addresses Gap (1)). I am currently working on developing a similar multi-modal dataset for auditing VLMs. I have also developed a gender inclusive text-based gender analyzer that predicts for the non-binary gender group (Jaiswal, Verma, and Mukherjee 2023) along with males and females.

RQ2b (developing model-centric bias mitigation algorithms) and **RQ3c** (auditing VLMs) are the current focus of my work and I plan to make further developments to both these topics as described follows.

Current Focus & Future Directions

(A) I have developed two adversarial audit strategies for FRSs and I am working towards developing an explainable audit strategy that can allow researchers to understand the model's preference for certain facial regions while performing the task of gender classification or face verification, allowing us to address **RQ1**.

(B) Developing data agnostic model-centric bias mitigation solutions is a significant challenge as the architectural changes can be sensitive and co-dependent. Moreover, these changes must be robust to introduction of new demographics as well adversarial variants. I plan to work towards addressing all these factors in my current work. This will address the challenges highlighted in **RQ2b**.

(C) Researchers have already audited VLMs for various social biases through well-designed audits and red-teaming exercises. I plan to audit VLMs for discrimination against non-binary individuals. I am currently developing multimodal datasets with non-binary representation. This will help me address **RQ3c**.

References

Jaiswal, S.; Duggirala, K.; Dash, A.; and Mukherjee, A. 2022. Two-face: Adversarial audit of commercial face recognition systems. In *AAAI ICWSM*.

Jaiswal, S.; Ganai, A.; Dash, A.; Ghosh, S.; and Mukherjee, A. 2024. Breaking the Global North Stereotype: A Global South-centric Benchmark Dataset for Auditing and Mitigating Biases in Facial Recognition Systems. In *AAAI/ACM Conference on AI, Ethics, and Society*.

Jaiswal, S.; and Mukherjee, A. 2022. Marching with the Pink Parade: Evaluating Visual Search Recommendations for Non-binary Clothing Items. In *ACM CHI EA*.

Jaiswal, S.; Verma, A. K.; and Mukherjee, A. 2023. Auditing gender analyzers on text data. In *IEEE/ACM ASONAM*.

Jaiswal, S. D.; Verma, A. K.; and Mukherjee, A. 2024. Mask-up: Investigating Biases in Face Re-identification for Masked Faces. *arXiv preprint arXiv:2402.13771*.