

BSTR CT

Human-facing AI platforms that users are compelled to interact with *involuntarily*, due to the absence of meaningful alternatives or lack of awareness, have become ubiquitous in modern digital infrastructures. These platforms are found in applications ranging from security surveillance to personalized recommendations. However, these technologies are not deployed equitably, leading to harmful and discriminatory outcomes. A significant limitation of currently deployed platforms is their inherent bias toward normative demographic groups, leading to disparate outcomes for other socially salient groups. Historically, bias evaluation and mitigation has also been restricted to the normative groups, thus further eroding the discourse on Responsible AI for socially salient/disadvantaged groups. Motivated by this, in this dissertation, we focus on developing frameworks for *auditing* commercial and open-source human-facing AI platforms for bias against socially salient groups and *mitigating* these disparities through novel datasets.

In the first part of this dissertation, we focus on benchmarking commercial and open-source Face Recognition Systems. A “black-box” audit of commercial FRSs using realistic inputs that simulate real-world deployment conditions reveals large-scale biases against minority race and gender groups. Further, a “white-box” audit of open-source FRSs disentangles the impact of datasets, architectures and loss functions on the disparity-accuracy trade-offs in these platforms. We demonstrate how the same architecture can be biased against opposite binary gender groups in different datasets. Thus, audit studies need to be contextualized within the scope of deployment, and results should not be generalized.

In our second work, we address the lack of representative face datasets from Global South demographics through real and synthetic face datasets. We first curate FAR-Face, a benchmark dataset with more than 50% images from the Global South. We

audit multiple FRSs on this dataset and reveal poor performance for females in the Global South. FRSs trained on the FARFace dataset report lower disparities across multiple benchmark datasets, indicating generalizability. Next, we propose SEM-CLORA, a novel scalable framework that leverages Vision Language Models and Diffusion Models combined with a trajectory-aware hard negative contrastive loss to generate diverse, realistic face images. FRSs trained on the synthetic data simultaneously improve accuracy and reduce disparity for socially salient groups.

In our third and final contribution, we benchmark AI platforms for biases against non-binary genders, a demographic that has been largely ignored in traditional analyses. We curate non-binary centric datasets from the image and text domains to benchmark visual search enabled e-commerce platforms and text-based gender analyzers, respectively. Our analysis of the visual search reveals that these systems systematically erase non-binary identities by forcing irrelevant binary recommendations. Similarly, existing textual systems misclassify non-binary users as binary female. To mitigate this, we propose mitigation strategies that fine-tune a BERT-based classifier on our textual social media dataset to significantly improve the performance.

Overall, this dissertation advances the field of Responsible AI by providing frameworks for auditing and mitigating biases in human-facing AI platforms. We focus on each component of a Responsible AI paradigm— inclusive datasets, third-party realistic audits and bias mitigation algorithms. Thus, this dissertation offers techniques on making AI more responsible and accountable for the global population.

Keywords: Responsible AI, Face Recognition Systems, Fairness and Bias, Global South, Generative AI, Diffusion Models, Non-binary Gender Analysis, Algorithmic Auditing, Synthetic Data Generation, Computer Vision.