Abstract

In recent years, Text-to-Image (TTI) diffusion models have witnessed a massive surge in popularity, and are now being used for a variety of applications across an ever-increasing number of professional domains, driven by their accessibility and remarkable advancements in generating high-quality and photo-realistic images from a simple textual prompt. However, these models are trained on datasets comprised of vast amounts of web data that are known to contain harmful biases and toxic content, and it has been proven that diffusion models trained on this data can learn and even amplify these biases. In this work, we conducted a comprehensive evaluation of emotional expressions on faces generated by two popular TTI diffusion models, DALL-E 2 and Stable Diffusion v 1.6, where we focused on how perceived facial emotions, race, and gender vary when changing the race and gender labels in the textual prompt given to the models. To detect facial expressions in synthetic images, we used Py-Feat, an open-source Python toolbox for facial expression analysis, and to detect race and gender, we used FairFace, an open-source gender, race, and age classification model. Through our evaluations, we found that both DALL-E 2 and Stable Diffusion exhibited numerous biases that ranged from confusing to harmful, including depicting faces as happy more often when 'woman' was in the prompt and negative when 'man' was,

as well as depicting faces as negative more often when 'Indian' was included. Additionally, Stable Diffusion was found to depict faces as negative much less often when 'White' was in the prompt compared to other race labels. The biases revealed in this work underscore the urgent need for rigorous testing and regular evaluations of popular and accessible TTI diffusion models, as understanding and mitigating the biases they exhibit is critical for ensuring they promote inclusivity rather than propagating existing social biases, which is instrumental for developing fair and equitable AI systems.