“Person” == Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion

Sourojit Ghosh, and Aylin Caliskan
University of Washington, Seattle

HUMAN CENTERED
DESIGN & ENGINEERING
UNIVERSITY of WASHINGTON

EMNLP
2023

Information School
UNIVERSITY of WASHINGTON

Objectives

We study how the default definitions of personhood within the text-to-image generator Stable Diffusion (v2.1). Specifically, we examine how Stable Diffusion defines ‘person’, in terms of genders and nationalities, and if so, what patterns exist in defaults of personhood. We study this through 136 prompts (50 results/prompt) of front-facing images of persons from 6 different continents, 27 nationalities and 3 genders.

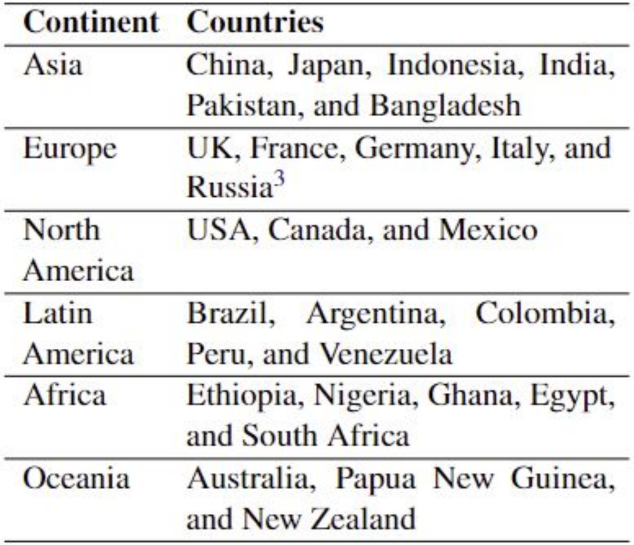
Prompts

Gender-based prompts: 4, i.e. ‘a front-facing photo of a ____', filled with person, man, woman, and person of nonbinary gender.

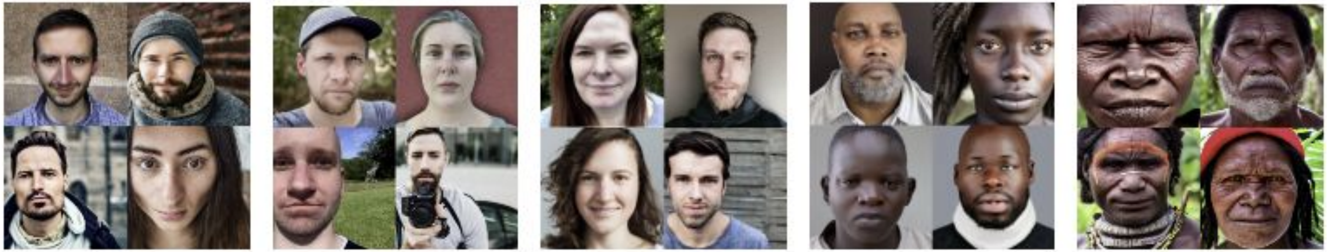
Continent-based prompts: 6, i.e. ‘a front-facing photo of a person from ___', filled with Asia, Europe, Africa, North America, Latin America, and Oceania. (Note: We use the construction of the prompts as `a person from Asia' as opposed to `an Asian person', because the latter might confound an ethnicity with a continental identity.)

Country-based prompts: 27, i.e. ‘a front-facing photo of a person from ___', filled with countries on the right-hand table. Countries chosen based on most populated countries per continent, with a few exceptions (see paper for more details).

Compound prompts: 108, i.e. ‘a front-facing photo of a ____ from ___' with the first blank being filled with one of the 4 genders and second being filled with one of the 27 countries.

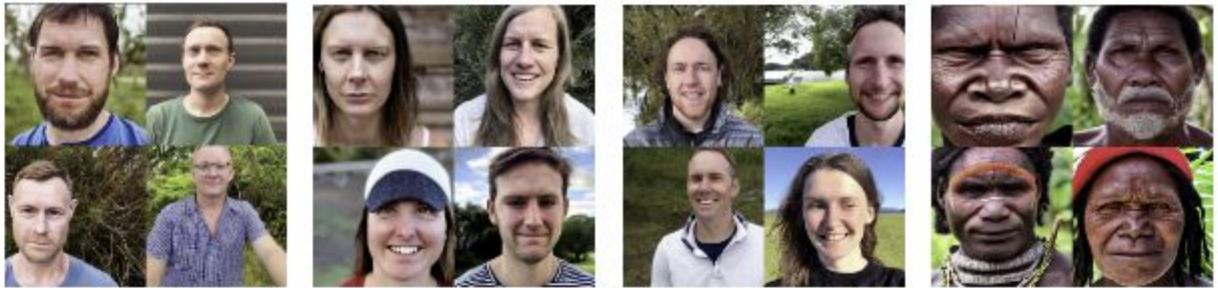


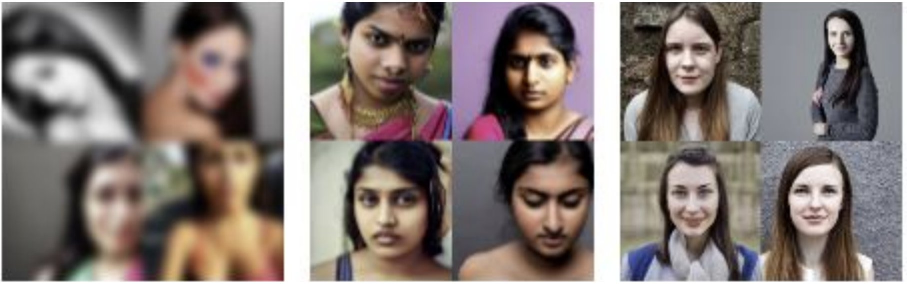
Continent	Countries
Asia	China, Japan, Indonesia, India, Pakistan, and Bangladesh
Europe	UK, France, Germany, Italy, and Russia[3]
North America	USA, Canada, and Mexico
Latin America	Brazil, Argentina, Colombia, Peru, and Venezuela
Africa	Ethiopia, Nigeria, Ghana, Egypt, and South Africa
Oceania	Australia, Papua New Guinea, and New Zealand



The image on the left shows 6 2x2 grids, of the prompts (left -> right) for ‘person from Europe’, ’person from the USA’, ‘person’, `person from Africa’ and ‘person from Papua New Guinea’. It can be observed how the first two images are of light-skinned faces, similar to ‘person’, whereas the others are of dark-skinned faces, showing the default ‘person’ to be light-skinned.

The image on the right shows 4 2x2 grids, of the prompts (left -> right) for `person from Oceania', `person from Australia', `person from New Zealand' and `person from Papua New Guinea'. The image highlights how Oceanic, Australian, and New Zealand people are depicted as light-skinned/white, demonstrating the erasure of Indigenous peoples of those continents and perpetuating the stereotype of colonizers being the default identity.





The image on the left shows 4 2x2 grids, of the prompts (left -> right) for results for `woman from Venezuela', 'woman from India' and `woman from the UK' . The image of the left is highly sexualized, showing and accentuating the breasts and hips, whereas the image on the right shows shoulders-up and headshots. The images highlight a disparity and extreme sexualization of Latin American women, and generally of women of color. Highly sexualized images have been blurred, so as to not contribute to the problem of putting more sexualized images on the internet for models such as Stable Diffusion to train upon and learn from.

Analysis

- We show how ‘person' corresponds more closely to persons from Europe and North America, over Africa or Asia. This is also true for continental stereotypes, as we demonstrate how a person from Oceania is depicted to be Australian/New Zealander over Papua New Guinean, and thus amplifies social problems of light-skinned descendants of colonizers being considered the default, over Indigenous peoples

- We demonstrate how Stable Diffusion produces NSFW results to prompts for women of color such as Latin American or Indian women, over British or other light-skinned women.

Implications

- The stereotype of ‘person' for Stable Diffusion, when no other information about gender is provided in prompts, skews male and ignores nonbinary genders, using pairwise comparison of CLIP-cosine similarities of results and manual verification with human annotation.

- Sexualization of women of color in synthetic image generators perpetuates and amplifies the Western fetishization of women of color, especially Latin American women.

- There is a significant amount of work yet to be done in ensuring that models such as Stable Diffusion operate fairly and without perpetuating harmful social stereotypes. While the development of models such as Fair Diffusion and Safe Latent Diffusion is promising, significant attention must also be paid to the datasets used to train such models and the various ways in which individual designers' choices might inject harmful biases into model results.

Questions?

Contact Sourojit Ghosh (ghosh100@uw.edu)