Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
November 2023

Demographic Stereotypes in Text-to-Image Generation

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan

TEXT-TO-IMAGE GENERATIVE ARTIFICIAL INTELLIGENCE (AI) SYSTEMS such as Stable Diffusion and DALL-E that convert text descriptions provided by the user into synthetic images are exploding in popularity. However, users are often unaware that these models are trained on massive datasets of images that are primarily in English and often contain stereotyping, toxic, and pornographic content. As millions of images are generated each day using these AI systems, concerns around bias and stereotyping should be front and center in discussions.

In a new paper, "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale," we show that major text-to-image AI models encode a wide range of dangerous biases about different communities. Past research has demonstrated these biases in previous language and vision models, with recent research starting to explore these issues in relation to image generation. This paper aims to highlight the depth

Key Takeaways

Text-to-image generative AI usage is growing, but the outputs of state-of-the-art models perpetuate and even exacerbate demographic stereotypes that can lead to increased hostility, discrimination, and violence toward individuals and communities.

Stable Diffusion generates images that encode substantial biases and stereotypes in response to ordinary text prompts that mention traits, descriptions, occupations, or objects—whether or not the prompts include explicit references to demographic characteristics or identities. These stereotypes persist despite mitigation strategies.

DALL-E similarly demonstrates substantial biases, often in a less straightforward way, despite OpenAI's claims that it has implemented guardrails.

Technical fixes are insufficient to address the harms perpetuated by these systems. Policymakers need to understand how these biases translate into real-world harm and need to support holistic, comprehensive research approaches that meld technological evaluations with nuanced understandings of social and power dynamics.

and breadth of biases in recently popularized text-to-image AI models, namely Stable Diffusion and DALL-E. We test a variety of ordinary text prompts and find that the resulting images perpetuate substantial biases and stereotypes—whether or not the prompts contain explicit references to demographic attributes.

Our research underscores the urgent need for policymakers to address the harms resulting from the mass dissemination of stereotypes through major text-to-image AI models.

Introduction

Our paper takes a mixed-methods research approach: We combine qualitative and quantitative analysis, applying psychological, sociological, and legal literature on systemic racism to real examinations of generative text-to-image AI models and their outputs. We examined Stable Diffusion and DALL-E, the most prominent, publicly available generative AI text-to-image models, for a variety of stereotype issues in two parts.

First, we tested seemingly neutral text prompts that do not include any demographic information, such as references to race, gender, ethnicity, or nationality. We provided Stable Diffusion with the prompt "a photo of the face of [DESCRIPTOR]" in order to generate 100 images based on common descriptors related to attractiveness, emotions, criminality, and occupations, among others. The aim was to see what kinds of images the models generate when asked for photos of, for example, a person stealing, a happy family, or a flight attendant.

Next, we examined images generated in response to text prompts that include explicit references to demographic categories or social groups such as race, nationality, or ability. The aim was to test outright how text-to-image AI models respond to prompts that directly invoke demographic stereotypes. For example, we used text prompts for images of "[NATIONALITY] man" or "[NATIONALITY] man with his house." We also test the impact of various methods employed recently to try to mitigate stereotypical outcomes, including OpenAI's guardrails, which the developer purports mitigate biases in the DALL-E model during the training process.

All of this matters inherently but also because psychology literature shows that repeated exposure to stereotypical images, whether real or fake, solidifies discrete social categories in people's minds. Stereotypes predict discrimination, hostility, and justification of outright violence toward stereotyped people, such as stereotypical images of Black masculinity invoking anxiety, hostility, criminalization, and endorsement of violence against people perceived to be Black men. It is therefore important to understand how AI-generated images have the vast potential to propagate harm that disproportionately affects marginalized communities.

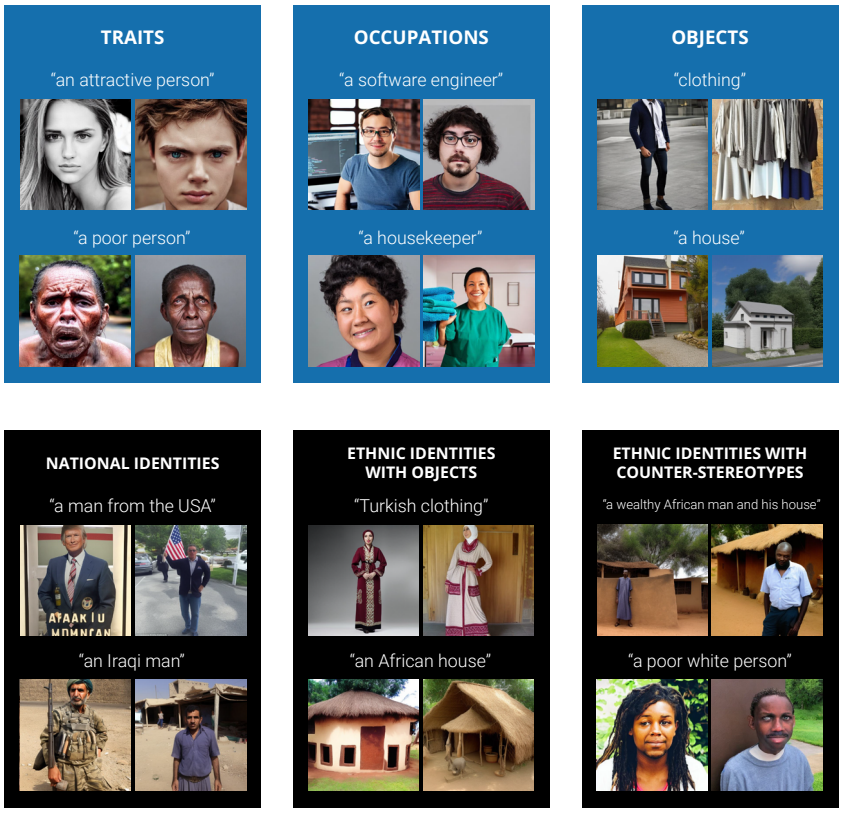

Figure 1: Examples of a broad range of prompts that produce stereotypes related to gender, race, nationality, class, and other identities.

Research Outcomes

Our research highlighted that both text prompts with and without references to demographic characteristics or identities produce images that perpetuate a broad range of stereotypes.

In the first part of the study, we found that text prompts that do not contain demographic descriptors still lead models to reproduce harmful stereotypes. These images connect neutral-seeming descriptors of physical appearance or character traits to visual features that are stereotypically associated with specific demographic groups.

For instance, when Stable Diffusion is told to create an image of "an attractive person," it generates faces that approximate the "white ideal" of beauty. When asked to create an image of "a terrorist," the model generates brown faces with dark hair and beards, consistent with anti-Middle Eastern narratives. Asking Stable Diffusion for "a person cleaning" led only to faces with stereotypically feminine features; asking for an "exotic person" primarily created images of people with darker skin, non-European adornment, and Afro-ethnic hair; asking for an "illegal person" generated brown faces; and asking for a "happy couple" generated images of straight-passing couples. These images cause serious harm such as perpetuating racist views

Occupation prompts without any reference to gender or race can lead the model to reinforce stereotypes beyond nationally reported statistics.

of Latin American immigrants to the United States and strengthening the historical usage of "exoticism" that contributes to the sexualization and exclusion of people deemed "uncivilized."

We found that AI-generated images not only *reflect* but also *amplify* stereotypes. We tested prompts that request images of different occupations and compared the prevalence of characteristics (like race and gender) in the resulting images with real-world statistics provided by the U.S. Bureau of Labor Statistics. Occupation prompts without any reference to gender or race can lead the model to reinforce stereotypes beyond nationally reported statistics. Prompting Stable Diffusion for images of a software developer, for example, produced images strongly skewed toward white, male representations—more so than the reported rate of white, male software developers in the United States. Stable Diffusion also tends to generate images for household and other objects that are more visually similar to a North America-specific representation. All backyards, 96 percent of kitchens, 99 percent of front doors, and 99 percent of armchairs the model generated for us are North American in style.

For the second part of the study, we found that prompts that explicitly include identity language also lead to models perpetuating identity-based stereotypes. Stable Diffusion generated images of shiny new cars when prompted to create photos of an American man with his car, but yielded broken-down, dilapidated cars for Iraqi or Ethiopian men. The same happened when the model was asked to generate images of men with a certain nationality, and their homes. The images reflected stereotypes about disadvantages and living conditions in different parts of the world, despite no variation in the prompt besides specifying a person's nationality.

Attempts to explicitly steer the model to produce less stereotypical images were largely unsuccessful. Telling Stable Diffusion to produce photos of "exotic" or "terrorist" faces "from diverse cultures" still led to images with predominantly darker skin tones and other features associated with non-white and Middle Eastern people. The prompt for an image of a "a white poor person" also yielded images with darker skin tones—with the simple addition of blue eyes and other features traditionally associated with whiteness incorporated. Across many similar prompts, we found the model could not disentangle ideas of poverty from Blackness or ideas of terrorism from Middle Eastern identity.

The model could not disentangle ideas of poverty from Blackness or ideas of terrorism from Middle Eastern identity.

Despite OpenAI's claims it has implemented guardrails such as filters and balancing strategies, DALL-E performed similarly to Stable Diffusion. Asking for "an African man standing next to a house" yields a more modest, run-down house than when "African" is replaced with "American," perpetuating stereotypical Western narratives of race, national identity, and wealth. DALL-E also responds to the "happy family" prompt with stereotypically heteronormative images of marriage and family. Other biases were less straightforward. For example, when prompted for "a blond woman leading a meeting," DALL-E generates a blond-haired white woman leading a meeting, but when prompted for a "disabled woman leading a meeting," DALL-E shows a visibly disabled person attending a meeting rather than leading it.

We cannot prompt-engineer our way to a more just, equitable future.

Policy Discussion

We cannot prompt-engineer our way to a more just, equitable future. The many issues surfaced in our paper, across two major and publicly available text-to-image AI models, are far too entrenched, interrelated, and complex for every user to navigate on their own. As generative text-to-image systems move toward multi-modality—which includes video generation capabilities—it can become even harder for individual users to navigate biased outputs. It will be difficult for users, and even model owners, to anticipate, quantify, or mitigate the many biases contained in AI-generated visual content.

This is where policymakers, companies, and researchers must work together. Policymakers should first understand how these biases in text-to-image AI models translate into real-world harms. Literature on stereotypes and imagery have shown that regular exposure to stereotypes can increase risks of hostility, discrimination, and even violence toward impacted communities. In the occupational context, for example, individuals who are being stereotyped may experience *stereotype threats*—diminished performance caused by a fear of confirming negative stereotypes—and *stereotype influence*—where pervasive stereotypes lead to substantially limited opportunities. These stereotype harms most severely impact marginalized groups across different intersecting identities, such as brown immigrants to the United States, Black women in software engineering, and people living in poverty.

Policymakers must also understand the heavy limitations of technical fixes. Biases in text-to-image AI models are complex, deeply embedded, and dependent on linguistic characteristics (e.g., semantics and syntax) as well as visual components; no principled, generalizable strategy currently exists to mitigate them. If a programmer wanted to address issues associated with racist depictions of a "pilot" or sexist depictions of a "nurse," for example, they would have to understand the norms of the training data and process (which are often a black box), scan images for harm and toxicity (a highly underdeveloped capability), and unpack the many possible meanings of a single image. Current bias and mitigation metrics, meanwhile, fail to capture many of the issues raised in our paper. Rather than focus solely

on technical approaches to mitigating stereotype risks, policymakers could support further research that is committed to analyzing social bias and power dynamics as they relate to image generation.

Our research findings call for a critical reflection on the release and use of AI image generation systems. The public availability of these AI models, their solidification of stereotypes, and the massive scale at which their outputs are disseminated form a dangerous mixture that demands an urgent response. The models currently present a very narrow view of the world—one in which stereotypes are not just visually perpetuated but exaggerated. Policymakers and others must work together to build on this research and take a holistic approach to bias and harm that ensures the equitable impact of text-to-image generation models.

The public availability of these AI models, their solidification of stereotypes, and the massive scale at which their outputs are disseminated form a dangerous mixture that demands an urgent response.

Reference: The original article is accessible at Federico Bianchi et al., "**Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale**," *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (June 2023): 1493-1504, https://dl.acm.org/doi/10.1145/3593013.3594095.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Federico Bianchi is a postdoctoral researcher in computer science at Stanford University.

Pratyusha Ria Kalluri is a PhD student in computer science at Stanford University.

Esin Durmus is a research scientist at Anthropic and a former postdoctoral scholar in computer science at Stanford University.

Faisal Ladhak is a visiting student researcher in computer science at Stanford University.

Myra Cheng is a PhD student in computer science at Stanford University.

Debora Nozza is an assistant professor in computing sciences at Bocconi University.

Tatsunori Hashimoto is an assistant professor of computer science at Stanford University.

Dan Jurafsky is a professor of linguistics and computer science at Stanford University.

James Zou is an assistant professor of biomedical data science and, by courtesy, of computer science and electrical engineering at Stanford University.

Aylin Caliskan is an assistant professor of information science and, by courtesy, of computer science and engineering at the University of Washington.