Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale

Federico Bianchi*
Stanford University
USA
fede@stanford.edu

Pratyusha Kalluri*
Stanford University
USA
pkalluri@stanford.edu

Esin Durmus*
Stanford University
USA
esdurmus@stanford.edu

Faisal Ladhak*
Columbia University,
Stanford University
USA
faisal@cs.columbia.edu

Myra Cheng*
Stanford University
USA
myra@cs.stanford.edu

Debora Nozza
Bocconi University
Italy
debora.nozza@unibocconi.it

Tatsunori Hashimoto
Stanford University
USA
thashim@stanford.edu

Dan Jurafsky†
Stanford University
USA
jurafsky@stanford.edu

James Zou†
Stanford University
USA
jamesz@stanford.edu

Aylin Caliskan†
University of Washington
USA
aylin@uw.edu

ABSTRACT

Machine learning models that convert user-written text descriptions into images are now widely available online and used by millions of users to generate millions of images a day. We investigate the potential for these models to amplify dangerous and complex stereotypes. We find a broad range of ordinary prompts produce stereotypes, including prompts simply mentioning traits, descriptors, occupations, or objects. For example, we find cases of prompting for basic traits or social roles resulting in images reinforcing whiteness as ideal, prompting for occupations resulting in amplification of racial and gender disparities, and prompting for objects resulting in reification of American norms. Stereotypes are present regardless of whether prompts explicitly mention identity and demographic language or avoid such language. Moreover, stereotypes persist despite mitigation strategies; neither user attempts to counter stereotypes by requesting images with specific counter-stereotypes nor institutional attempts to add system “guardrails” have prevented the perpetuation of stereotypes. Our analysis justifies concerns regarding the impacts of today’s models, presenting striking exemplars, and connecting these findings with deep insights into harms drawn from social scientific and humanist disciplines. This work contributes to the effort to shed light on the uniquely complex biases in language-vision models and demonstrates the ways that the mass deployment of text-to-image generation models results in mass dissemination of stereotypes and resulting harms.

ACM Reference Format:
Federico Bianchi*, Pratyusha Kalluri*, Esin Durmus*, Faisal Ladhak*, Myra Cheng*, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky†, James Zou†, and Aylin Caliskan†. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’23), June 12–15, 2023, Chicago, IL, USA. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3593013.3594095

*These authors contributed equally to the realization of this project.
†Corresponding authors: jurafsky@stanford.edu, jamesz@stanford.edu, aylin@uw.edu

Content warning: This paper includes and discusses model-generated images that may be offensive or upsetting.

1 INTRODUCTION

There has been a rapid rise of machine learning models able to convert user-written text descriptions into images, with several of these models now available for anyone online to use. These models — of which Stable Diffusion [14, 59] and Dall·E [57] are the most popular — often require little to no prior technical expertise and can be used to generate thousands of images in a few hours. Industry publicization, hype, and ease of access have already led to millions of users, generating millions of images a day [52]. Moreover, these
A large, long-standing body of psychology literature shows when online and thus propagated to many downstream outputs via the biases encoded in image generation models that are easily available in these models proliferating at a massive scale in the millions of generated images. We focus on the prototypical, publicly available Stable Diffusion model by [59], as all components of the model are openly documented and available for analysis. We also investigate harmful representations of demographic groups in these kinds of models despite user interventions (careful prompting) and institutional interventions (e.g., the so-called system “guardrails” added to Dall-E). This work is grounded in a mixed-methods research orientation. In all studies, we aim to foreground striking cases of stereotype-inducing prompts, exemplar images, and rich, qualitative analysis that draws out connections from these prompts and images to psychological, sociological, and critical race theory literature on the particular, discovered biases and their consequences. In drawing out these connections, we have aimed to strike a balance between using language accessible to a wide audience (including when discussing extraordinarily complex topics like race and gender), while doing key translational work surfacing rich and deep insights from these social disciplines. When appropriate, we also include supplementary quantitative analysis; in particular, to illuminate the model’s internal representations. The foregrounding of striking exemplars and qualitative analysis is crucial to the aims of this work, for two key reasons. First, qualitative analysis is necessary when it is desired and valuable to explore, characterize, and demystify a space of phenomena that are not yet well-theorized [3, 5, 43], as is precisely the case with these large, newly emerging text-to-image models whose dynamics are far from well-understood and nonetheless rapidly proliferating. Second, this work is in part a response to that rapid proliferation and the urgently growing need to call for attention and intervention on the harms of these models. Our work seeks to leverage the significant research indicating that exemplars of social harms, and in particular, exemplifying imagery — often more so than only quantifying base rates of harms — constitute “a powerful means of creating risk consciousness and of motivating protective and corrective action” [78]. This is especially important to counterbalance the prevalence of generated image exemplars cherry-picked for their aesthetic or unproblematic qualities, including those featured heavily on the online sites where users go to use these models. On the basis of these motivations, this work characterizes the prevalence of dangerous racial, ethnic, gendered, class, and intersectional stereotypes across a wide range of natural-language prompts (Figure 1), illuminating these models’ vast potentials for propagating harm along many axes of demographic identity.

First, simple user prompts containing character traits and other descriptors generate images perpetuating stereotypes. We explore the outputs resulting from user prompts that contain common descriptors including character traits, occupations, and household items/objects. For example, an attractive person generates faces approximating a “White ideal” [32], perpetuating the history of subordinating persons who do not fit this ideal as lesser [42, 69], and a terrorist generates brown faces with dark hair and beards, consistent with narratives that has been used to rally for anti-Middle Eastern violence [15, 16, 26].

For descriptors that have comparable real-world statistics across demographic groups, such as occupations, we find cases of near-total stereotype amplification. In these cases, the model does not merely reflect societal disparities, and instead actually exacerbates them. For example, 99% of the generated software developer

Figure 1: A broad range of prompts produce stereotypes related to gender, race, nationality, class, and other identities. Complex biases persist for prompts that do not use identity language (top row), prompts that mention identities (bottom row), and prompts that include explicit countering of stereotypes (bottom row, middle). We present two random examples for each prompt.
images are represented as white according to a pre-trained model, while in the country where the foundational training dataset was constructed (the U.S.), only 56% of software developers identified as white.

Furthermore, when a prompt mentions social groups (e.g. race or nationality), Stable Diffusion generates images that tie specific groups to negative or taboo associations like malnourishment, poverty, and subordination. For everyday things like household objects, we find that the model implicitly makes similar associations: the image of an Ethiopian man and his car produces an image of poverty, while the same prompt with American does not. The model also perpetuates cultural defaults and harmful norms for various settings, from everyday events to special occasions: an image of a front door is perceived as if that the door is in North America, while happy couple generates only straight-passing couples. More concerning, even when these stereotypes are explicitly countered in the prompts (such as adding wealthy to a prompt that otherwise generates unintended images of poverty), we demonstrate that the model may still be unable to generate these images at all and continue to perpetuate stereotypes. Despite the veneer of claiming to generate anything imaginable, Stable Diffusion is actually limited to generating images that align with dominant stereotypes, even when asked to produce the contrary. We find that these associations are mitigated by neither carefully written user prompts nor the “guardrails” against stereotyping that have been added to models like Dall-E [51].

We demonstrate these issues using simple natural-language prompts, meaning the patterns that we identify are easily accessible and plausible occurrences and are thus cause for serious concern in their potential prevalence. We discuss the challenges of mitigation due to the countless dimensions and intersections of social groups and the compounding nature of language-vision biases. This paper does not aim to survey model strengths, quantitatively assess all possible mitigation strategies, or decisively identify the optimal mitigation strategy or broader solution going forward; it is in fact impossible to do so, in part because, everyday, new changes to the development and deployment of these models are cropping up, and being proposed, explored, tested, or implemented, often outside the purview of the public. Rather, this paper takes the role of drawing attention to concerns regarding the impacts of today’s models, presenting striking evidence, and connecting these findings with deep insights into harms that are richly discussed in social scientific and humanist disciplines. The accessibility of these models, combined with the extent to which they reify social categories and stereotypes, form a dangerous mixture, as use cases for these models, including creating stock photos [40] or supporting creative tasks [50], render these issues particularly troubling. These applications are mass disseminating images and stereotypes while failing to articulate and invisibilizing other ways of being. As these models create biased and harmful snapshots of our world in data, media, and art, our work calls for a critical reflection on the release and use of image generation systems and AI systems at large. We release additional information on an online repository.1

2 PROMPTS WITH NO IDENTITY LANGUAGE PERPETUATE AND AMPLIFY STEREOTYPES

One belief may be that by Constraining prompts to seemingly neutral language or language that avoids identity descriptors, the territory of stereotypes and biases is also avoided. This notion is related to the ideology of “colorblindness,” which has long been criticized for perpetuating racism [7, 71]. In this section, we explore a variety of harmful stereotypes that arise from prompts that do not mention any identity or demographic group at all. We view the striking outputs of these prompts as a visceral demonstration that users not referencing race, ethnicity, or gender language are now capable of unintentionally mass generating and disseminating images perpetuating historically dangerous stereotypes.

2.1 Human traits and descriptors: Perpetuating stereotypes

We begin by investigating this question: can simple descriptions that do not reference race, gender, ethnicity, or nationality language nonetheless lead models to reproduce harmful stereotypes? We present ten cases confirming that the answer is unequivocally yes. For each of ten commonly-used human descriptors, the prompt “A photo of the face of [DESCRIPTOR]” (e.g. “A photo of the face of an attractive person”) was fed to Stable Diffusion to generate 100 images. Descriptors and a random sample of the generated images are presented in Figure 2.2

We find that the generated images reify many dangerous social- etal associations by tying descriptors to visual features that are stereotypically associated with specific socially-constructed demographic groups [30]. The Stable Diffusion model defines attractiveness as near the “White ideal” (blue eyes, pale skin, or long, straight hair; [32]) and ties emotionality and seductiveness specifically to stereotypically white feminine features. Relatedly, we found that a person cleaning generates only faces with stereotypically feminine features. Meanwhile, the model exoticizes people with darker skin tone, non-European adornment, and Afro-ethnic hair [65]. The former positions femininity as subordinate relative to white masculinity [8], while simultaneously perpetuating the legacy of whiteness as the default ideal, further subordinating those who do not belong to the white monolith [42, 69]. The term “exotic” has a long history of being used to refer to populations that have been deemed “uncivilized” by a dominant group [47] and continues to contribute to sexualization and exclusion [46].

Further, we find a poor person and a thug generate faces with dark skin tone and features stereotypically associated with Blackness [35], and a person stealing similarly generates faces with dark skin and stereotypically Black features, perpetuating patterns in media that are known to invoke anxiety, hostile behavior, criminalization, and increased endorsement of violence against people perceived as Black men [9, 25, 49, 63]. Prompting the Stable Diffusion model to generate a terrorist results in brown faces with dark hair and beards, consistent with the American narrative that terrorists are brown bearded Middle Eastern men, justifying bans and violent policies against persons perceived as in this group [15, 16, 26]. Similarly,

---

1https://github.com/vinid/text-to-image-bias

2Other prompt templates resulted in similar results, and these are presented in the Appendix; for those interested in furthering this line of investigation, all generated images are available upon request.
an attractive person generates brown faces, mirroring the American concept of “illegal” Latin American immigrants [11, 21].

Broadly, we find that these outputs perpetuate stereotypes by entangling stereotypical features of demographic groups with neutral-seeming descriptors. Note that in some cases, the task of producing an image in response to a prompt is harmful in and of itself. For instance, the very notion of generating an image of “the face of a poor person” is problematic in and of itself, as race, class, and other social categories are not immutable [61], and more broadly, images of particular characteristics that are or are meant to be uncorrelated with visual attributes cannot be generated without making dangerous assumptions.

Yet another dimension of bias is revealed in the models’ generations when prompted with descriptors of social structures regarding groups of people. For prompts of “a happy couple” and “a happy family,” the straight-passing output images reinforce heteronormative social institutions, which presume that marriage and family structures are based on different-sex couples [36, 37]. These normative assumptions alienate those who do not conform to these norms, contributing to the well-documented phenomenon of minority stress: those with LGBTQ+ identities disproportionately experience stress and other mental health consequences as a result of homogenizing stereotypes, stigma, and discrimination [19, 45].
2.2 Occupations: Stereotype amplification

Given the many cases of model-generated images perpetuating stereotypes, we next turn our attention to quantifying the potential for stereotype amplification. Stereotype amplification is the process of real-world correlations between social identities like race and gender being distorted and exaggerated, possibly to the point of being perceived as ubiquitous [55]. Prior work has demonstrated that previous language models and word embeddings can amplify biases in general, and stereotypes in particular, beyond rates found in the training data or the real world [23, 77].

Given the foundational training dataset was constructed in the U.S., and given observations that machine learning representations reflect American norms and values and reproduce inequities of American society (Section 2.3, [31, 72, 73]), we focus on quantifying the extent of amplification of U.S. social associations. Further, we focus on the association between racial and gender categories and occupation, because in the U.S., race and gender are pervasively conceptualized as core demographic categories used socially and recorded by the census bureau, and national surveys quantify occupation demographics in terms of these categories [66]. We are interested in the extent to which the U.S. ‘official’ demographic categorizations (Male/Female, White/Black/Asian) and associated occupations are perpetuated in the Stable Diffusion model and generated images. (We further describe our use and the broader social context of these social categories in the Appendix.) For example, historical forces have shaped who becomes software developers, and this group reportedly self-identifies as majority white men. Given a prompt referring to a software developer, does the model lessen this skew and generate diverse skin tones and features, reflect this skew, or amplify this skew, representing software developers with nearly exclusively stereotypically white male features? We uncover many instances of the latter: near-total stereotype amplification.

Based on the availability of data from the U.S. Bureau of Labor Statistics, we present ten occupations, of which several have been studied in prior work on biases in natural images, e.g. software developer and housekeeper [33, 68, 76], and others reportedly have substantial demographic imbalances [66]. First, we generate images of each occupation, and then we analyze the way the model represents these images: For each occupation, the prompt “A photo of the face of [OCCUPATION]” (e.g. “A photo of the face of a housekeeper”) was fed to the model, the model was used to generate 100
Figure 4: Quantifying stereotype amplification. For each occupation, we compare the reported percent of the occupation that self-identified as female and non-white (from U.S. Bureau of Labor Statistics 2021 data) to the percent of the occupation-generated images the model represented as female and non-white. In many cases, gender imbalance in an occupation corresponds to extreme gender imbalance in the generated images, e.g. a slight majority of flight attendants reportedly identified as female, but 100% of flight attendant images were perceived by female according to the model-based approach outlined above. Regardless of occupation demographics, the model represents several of the most prestigious, high-paying professions like software developer and pilot as white.
Algorithmic amplification of associations between gender and race poorly on images from low-income countries. Using a set of objects, household objects, such as beds, doors, etc. and find that they work ble Diffusion encode American norms. In prior work, De Vries et al. racial hierarchies, we explore whether generated images from Stages of everyday objects. Building upon the findings of Wolfe et al. that stereotypes and norms are also injected into generated im-

groups and other problematic biases. In reality, however, we find entities, we avoid reproducing representations of demographic 

One might hope that by generating images solely of non-human entities, we avoid reproducing representations of demographic groups and other problematic biases. In reality, however, we find that stereotypes and norms are also injected into generated images of everyday objects. Building upon the findings of Wolfe et al. that visual semantic models like CLIP reproduce American racial hierarchies, we explore whether generated images from Stable Diffusion encode American norms. In prior work, De Vries et al. test object recognition systems on a dataset of 117 classes of household objects, such as beds, doors, etc. and find that they work poorly on images from low-income countries. Using a set of objects from this list, for each object, we used Stable Diffusion to generate 100 images using the following prompts: (1) A general prompt (“a photo of [OBJECT]”) (2) A North-America-specific prompt (“a photo of [OBJECT] in North America”) (3) An Asia-specific prompt (“a photo of [OBJECT] in Asia”) and (4) An Africa-specific prompt (“a photo of [OBJECT] in Africa”). We present random examples in Figure 5. We see that seemingly neutral prompts about objects produce decisively culture-specific images: the general prompt produces outputs that are most similar to the North-America-specific prompt, while visually differing most greatly from outputs with the Africa-specific prompt.

To quantify this finding, we employ a strategy similar to that used in Section 2.2. We again study the representations in CLIP [56], the core representational component of Stable Diffusion. For each of the continent-specific prompts (e.g., the North-America-specific prompt), we identify an archetypal vector representation of household objects from this continent by feeding the prompt-generated images to CLIP’s image encoder to generate vector representations, and we average them – thus obtaining a single archetypal vector representation for each continent (i.e., a vector for North America, a vector for Asia, and a vector for Africa). We can then document, for each general-prompt-produced image, which continent vector the general-prompt-produced image is closest to (in cosine distance). In this way, we identified the continent prompt for which the general prompt with no continent specifier yields the most similar results. Results confirm what we see with visual inspection: for example, for 100% of backyards, 96% of kitchens, 99% of front doors, and 99% of armchairs, the general-prompt-produced image is more similar to the North America-specific representation than the representation of any other continent.

This finding connects to the American-focused demographics and norms of Internet-based datasets [4]. Notably, this does not reflect real-world population statistics: based on population, there are many more front doors and kitchens in other parts of the world, yet the generated outputs only reflect American ones. By generating images that are stylized as American when prompted with
everyday objects, these models create a version of the world that further entrenches the view of American as default. This “view from nowhere,” i.e. hiding a specific perspective and set of assumptions under the guise of neutrality, has been long-studied and criticized by sociologists for contributing to the exclusion and ostracizing of those who do not belong to the default group [27, 58].

3 PROMPTS WITH IDENTITY LANGUAGE PERPETUATE STEREOTYPES, DESPITE MITIGATION EFFORTS

In the previous section, we demonstrate how prompts that do not use identity language exhibit and perpetuate stereotypes. In this section, we find that prompts that contain explicit references to identity or demographic categories produce images imbued with many layers of identity-based stereotypes. Furthermore, stereotypes remain in the images even when prompts explicitly request counter-stereotypical images.

3.1 Stereotyping representations of groups

We find that, when prompting with identity language, a wide spectrum of visual components of the generated images—from the people to the objects to the background—can all reinforce systemic disadvantages. We present a variety of striking examples, many unique in how they encode systemic disadvantage and bias, but consistent in doing so. In Figure 6, we present examples of how signals of disadvantage or bias are reproduced through socially loaded visual components. For each prompt, we present two random examples. Comparing the generated images of “a photo of [NATIONALITY] man with his car,” it is apparent that the car in the image with the American man is shiny and new, while the car in the picture with the other nationalities are broken and in bad condition, despite this difference not being in any way stated in the prompt.

Examining the outputs of “[NATIONALITY] man” or “[NATIONALITY] man with his house” produces similar patterns. These patterns reinforce the narrative that African countries like Ethiopia are defined by poverty, while individuals from Middle-Eastern countries like Iraq cannot be defined apart from their involvement with war. These cases demonstrate the complexity of these biased associations and how identity terms in language are reflected to vision in many nuanced ways. This effect is also apparent in our generations of images of household objects, which is described in Section 2.3. As displayed in Figure 5, we find that these objects reflect the same disadvantaging patterns and stereotypes about different continents. These examples touch the surface of the multitude and perverseness of stereotypes being perpetuated, which deserve to be deeply explored and systematically assessed in their own right.

Figure 6: Examples of complex biases in the Stable Diffusion model. The generated images of Ethiopian and Iraqi cars are in worse condition than that of the American without any explicit prompting. The model broadly encodes ethnic stereotypes: the prompt “an Ethiopian man” often generates images of apparently malnourished individuals, while “an Iraqi man” can generate images related to war and military force. The prompts are written above the corresponding generated images. We present two random examples for each prompt.

3.2 Stereotypes despite counter-stereotypes

One strategy to mitigate stereotypes and steer the model toward generating less harmful outcomes is to add targeted modifiers to prompts. We show that in many cases, this does not eliminate stereotypes in the generations of Stable Diffusion. Moreover, we show that even given explicit modifiers that mention identities that counter stereotypes, the biases persist in the generations.

Recent work by Bansal et al. [2] shows that in some cases it is possible to modify prompts to get more diverse generations, e.g., by adding “from diverse cultures” at the end of the prompt to obtain images with more diverse cultural representation. Indeed, we find that...
prompts like "a photo of the face of [DESCRIPTOR] from diverse cultures" can in some cases force Stable Diffusion to generate more diverse images (e.g., *software developer, flight attendant*). However, in other cases character traits still show persistent stereotypical patterns even with this rewriting. We generated 100 images of "a photo of the face of [DESCRIPTOR] from diverse cultures" for the descriptors *exotic* and *terrorist* and find that the produced images continue to exhibit darker skin tones and other features associated with non-white and Middle-Eastern identities respectively. We hypothesize that this is because the phrase "diverse cultures" is interpreted by the model as cultures that are diverse relative to the American default of whiteness [22, 38, 54].

Moreover, we find that, even when prompts are carefully written to oppose the observed stereotype, socially salient stereotypes persist. For instance, to counter the pattern of "a poor person" generating faces with dark skin tone (Figure 2), a possible attempt could be using the prompt of "a white poor person." However, we find that even using this modified prompt, most of the images exhibit darker skin tones and merely incorporate some features that are typically associated with whiteness, such as blue eyes. In other words, the model continues to generate stereotypically Black faces for "poor person" despite being explicitly prompted to do otherwise. Similarly, to counter the dominant stereotype of "a terrorist," we attempt to use the prompt of "a white terrorist." With this modified prompt, we find that many of the generated images have long beards that are stereotypically associated with Middle-Eastern men (Figure 7). These examples suggest that the model is fundamentally unable to disentangle poverty from Blackness and terrorism from Middle-Eastern identity regardless of the text of the prompt.

Further, even when prompted with the description "a photo of an African man and his fancy house," which intentionally includes the modifier "fancy" to subvert inappropriate associations with poverty, Stable Diffusion generates an image that continues to reify the notion that an African man always lives in a simple hut or broken structure, in comparison to the American man (Figure 7). The situation is no better, when the prompt is "a photo of an African man and his mansion," as this prompt again reproduces the same association (Figure 7). Another cause for concern arises from another aspect of the mitigation attempt with prompt rewriting, in the image generated from the prompt "a photo of a wealthy African man and his house." While the house stays the same, the man now dons a suit—a Western status signal for wealth. In this way, the concepts of wealth and Western society continue to be conflated. Further, the Stable Diffusion model makes these errors even when such photos clearly exist on the web: even Google Image Search—which in the past has sparked controversy for its amplification of societal bias [34, 44, 48, 62]—is capable of showing us, upon searching for "African man and his mansion," an African man, dressed in opulent African-style clothing, in front of an ornate house. Stable Diffusion, then, is capable of exhibiting more stereotypes than what is deemed acceptable by the standards that govern the creation of stock photos. Thus, even when prompts are actively written to subvert existing societal hierarchies, image generation models often cannot reproduce these imaginations. This reflects the notion that colonial and power relations "can be maintained by good intentions and even good deeds", certainly by well-intentioned prompts[39].

### 4 STEREOTYPES ARE PERPETUATED DESPITE INSTITUTIONAL GUARDRAILS: THE CASE OF DALL·E

A fundamentally different mitigation approach, which centers model owners rather than users, is to actively implement "guardrails" to mitigate stereotypes [51]. When making DALL·E widely accessible, OpenAI attempts to mitigate biases by applying filtering and balancing strategies to improve the quality of the data used to train the model; it also has a mechanism to prevent the generation of images from prompts that are viewed as dangerous. The exact mechanisms of the "guardrails" are not fully disclosed by its creators, which further complicates the issue, as we do not know what domains they consider, let alone what notions of bias or fairness they use. Nevertheless, we show that complex, dangerous biases still exist in DALL·E. We present striking examples, all of which appeared on the first page of results from DALL·E upon feeding in the corresponding prompt.

First, we reproduce the experiment of generating images of everyday objects, as well as these objects in various continents (Section 2.3). Sample images, analogous to Figure 5, are in Figure 7. We find similar trends in the Stable Diffusion and DALL·E produced images, where not specifying any country results in images that are the most similar to "North America," while including "Africa" in the prompt leads to images that reinforce the narrative that African countries are first and foremost places of poverty.

Interestingly, due to DALL·E’s "guardrails" that ensure the generated images exhibit a diverse range of skin tones, we find that for prompts mentioning occupation, the DALL·E outputs have a range of skin tones and stereotypically masculine/feminine features. We posit that this is because occupational biases have been amongst the most well-studied in computer science bias scholarship [17] and thus have been explicitly considered by the creators of DALL·E. However, other subtler biases, biases more fully studied in other disciplinary contexts, as well as biases on other dimensions beyond gender and race, persist.

We find that many other stereotypes and patterns that plague Stable Diffusion occur with DALL·E (Figures 6 and 8). The prompt "An African man standing next to a house" produces images of houses that appear simpler and more worn-down compared to the images produced by replacing the word "African" with "American." Notably, when the model generates an African man and an American man simultaneously, with the prompt "a photo of an African man and an American man standing next to a house," an American house apparently in good condition is produced (Figure 8, first row). The model seems unable to disentangle the constructed concepts of race, nation, and wealth, reflecting the ways that these characteristics have been tied closely together in the past. This is deeply concerning: how can we dream and move beyond the racist hierarchies constructed by the West [20] if such images only become more widespread?

We present various additional examples of complex biases in DALL·E revealing that these pernicious hierarchies extend beyond race and wealth. These examples demonstrate the persistence of biases despite mitigation attempts via explicit counter-stereotypes and identifiers in the prompts. First, when prompted with "a disabled woman leading a meeting," the model does not produce an
image where the visibly disabled persons appear to be leading. Instead, they appear to be listening to someone else, who is evidently in a position of authority. This problem disappears when the word "disabled" is replaced with "blonde." That the model is not able to depict an intentionally crafted scenario, in which disabled women can lead meetings, underscores the ways it can deepen existing ableism (Figure 8, second row). Furthermore, like Stable Diffusion, the model’s outputs for “a happy couple” and “a happy family” reinforce heteronormative ideals (Figure 8, third and fourth row). Despite the implemented “guardrails,” Dall-E demonstrates substantial biases along many axes, often in less straightforward ways.

5 CONCLUSION

In this work, we demonstrate the presence of dangerous biases embedded in image generation models. Given these technologies are now widely available and generating millions of images a day, there is serious and, we illustrate, justified concern about how these AI systems are going to be used and how they are going to shape our world. It is likely to be challenging or impossible for users or model owners to anticipate, quantify, or mitigate all such biases, especially when they appear with the mere mention of social groups, descriptors, roles, or objects. This is in part due to the multifacetedness of social identity—there are countless axes and intersections of social groups [24]. The issues being surfaced here necessitate thinking beyond reductionist computational approaches and making long-term commitments to analysis of the evolving dynamics of social biases and power relations. The compounding issues in the multi-modal domain, as AI systems are headed towards increasing multi-modality (for example, generating videos), have only increasingly drastic impacts on our lives. Our analyses show that even better prompts, carefully curated to promote diversification and subvert undesired stereotypes, cannot solve the problem because images encode and display a multitude of information beyond the specifications of a prompt. We also cannot expect end users of these technologies to be careful as we have been when prompting for images. We cannot prompt-engineer our way to a more just, inclusive and equitable future.

There are several reasons why mitigating bias in image outputs from language-vision models like Stable Diffusion will be uniquely challenging. The generated output images necessarily contain many aspects that are not explicitly specified in the prompt. For instance, if the prompt references an object, the model must infer all of the characteristics of this object and cannot leave out information that has otherwise been unspecified. Thus, the output adheres to norms reflective of the training data and process. Images contain many more dimensions than text, offering seemingly endless opportunities to pack subtle meaning within. Alongside these opportunities is the serious harm, which we demonstrate in this paper, of propagating biases that are completely beyond the boundaries of the issues on which current bias metrics and mitigation efforts focus. Moreover, automated methods to scan text for harm and toxicity, even with significant shortcomings, are still far more well-developed than those to scan images [29]. Ultimately, since these biases are complex and dependent on both linguistic characteristics (semantics, syntax, frequency, affect, conceptual associations) and many components in the visual domain, thus far there exist no principled and generalizable mitigation strategy for mitigating such broadly and deeply embedded biases.

We urge users to exercise caution and refrain from using such image generation models in any applications that have downstream effects on the real-world, and we call for users, model-owners, and society at large to take a critical view of the consequences of these models. The examples and patterns we demonstrate make it clear that these models, while appearing to be unprecedentedly powerful and versatile in creating images of things that do not exist, are in reality brittle and extremely limited in the worlds they will create.
ACKNOWLEDGMENTS

This work was funded in part by the Hoffman–Yee Research Grants Program and the Stanford Institute for Human-Centered Artificial Intelligence. Additional funding comes from a SAIL Postdoc Fellowship to ED, an NSF CAREER Award to JZ, an NSF Graduate Research Fellowship (Grant DGE-2146755) and Stanford Knight-Hennessy Scholars graduate fellowship to MC, and funding from Open Philanthropy, including an Open Phil AI Fellowship to PK. This material is also based on research partially supported by the U.S. National Institute of Standards and Technology (NIST) Grant 60NANB20D212T. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of NIST.

REFERENCES

[1] David M. Amodio and Patricia G. Devine. 2006. Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behavior. Journal of personality and social psychology 91 4 (2006), 652–621.
[2] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kui-Wei Chang. 2022. How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? ArXiv preprint abs/2210.15230 (2022). https://arxiv.org/abs/2210.15230
[3] Howard S. Becker. 1996. The epistemology of qualitative research. Ethnography and human development: Context and meaning in social inquiry 27, 53-71 (1996).
[4] Emily M. Bender, Tummit Gehr, Angelina McMullan-Major, and Shmargaret Smittleh. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 610–623.
[5] Bruce Lawrence Berg. 2001. Qualitative research methods for the social sciences. Allyn & Bacon.
[6] Abeba Birhane, Vinay Udai Prabhu, and Emmanuel Kamlehwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. ArXiv preprint abs/2100.01963 (2021). https://arxiv.org/abs/2100.01963
[7] Eduardo Bonilla-Silva. 2006. Racism without racists: Color-blind racism and the persistence of racial inequality in the United States. Rowman & Littlefield Publish.
[8] Victoria L. Brescoll. 2016. Leading with their hearts? How gender stereotypes of emotion lead to biased evaluations of female leaders. Leadership Quarterly 27 (2016), 415–428.
[9] Diana J. Burgess, Yingzai Ding, Margaret K. Hargraves, Michelle van Ryn, and Sean M. Phelan. 2008. The Association between Perceived Discrimination and Underutilization of Needed Medical and Mental Health Care in a Multi-Ethnic Community Sample. Journal of Health Care for the Poor and Underserved 19 (2008), 894 – 911.
[10] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. In ECCV. 22, 4 (2004), 623–646.
[11] Kevin L. Keenan. 1996. Skin Tones and Physical Features of Blacks in Magazine Advertisements. Journal of Mass Communication Quarterly 73, 4 (1996), 905–912. https://doi.org/10.1177/002229139607300410
[12] Amanda E Lewis. 2004. What group?” Studying whites and whiteness in the era of color blindness. Social problems 45, 3 (2008), 292–306. https://doi.org/10.1177/0037093503258743
[13] Marie De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Chris vine Brescoll. 2016. Gender stereotypes of emotion and the privilege of partial perspective. In Feminist theory reader. Routledge. 303–310.
[14] Deborah Hallinan. 2019. Indirect discrimination and the duty to avoid compounding injustice. Foundations of Indirect Discrimination Law, Hart Publishing Company (2018), 2017–255.
[15] Philip Aitha Goff, Jennifer L. Eberhardt, Melissa J. Williams, and Matthew Jackson. 2008. Not yet human: implicit knowledge, historical dehumanization, and contemporary consequences. Journal of personality and social psychology 94 2 (2008), 292–306.
[16] Nandita Grewal. 2003. Transnational America: race, gender and citizenship after 9/11. Social Identities 9, 4 (2003), 535–561. https://doi.org/10.1080/135043303200174669
[17] Donna Haraway. 2020. Situated knowledges: The science question in feminism and the privilege of partial perspective. In Feminist theory reader. Routledge. 303–310.
[18] Abram Kardiner and Lionel Oversey. 1951. The mark of oppression: a psychosocial study of the American Negro.[1st ed. ed.]. Norton New York. xvii, 396 p. pages.
[19] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015. Bo Gege, Jinwoo Kim, Kori Inkpen, and Woontack Woo (Eds.). ACM, 3819–3828. https://doi.org/10.1145/2702123.2705202
[20] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd annual acm conference on human factors in computing systems. 3819–3828.
[21] Kevin L. Keenan. 1996. Skin Tones and Physical Features of Blacks in Magazine Advertisements. Journal of Mass Communication Quarterly 73, 4 (1996), 905–912. https://doi.org/10.1177/002229139607300410
[22] Celia Kitzinger. 2005. Heteronormativity in action: Reproducing the heterosexual nuclear family in after-hours medical calls. Social problems 52, 4 (2005), 477–498.
[23] Roger N Lancaster. 2003. We are the trouble with nature: Sex in science and popular culture. Univ of California Press.
[24] Amanda E Lewis. 2004. What group?” Studying whites and whiteness in the era of color blindness. Sociological theory 22, 4 (2004), 623–646.
[25] Max Liboiron. 2021. Pollution is colonialism. In Pollution Is Colonialism. Duke University Press.
[26] Natasha Lomas. 2022. shutterstock to integrate OpenAI’s DALL-E 2 and launch fund for creator artists – techcrunch.com. https://techcrunch.com/2022/10/25/shutterstock-openai-dalle-2-ai/. [Accessed 01-Nov-2022]
[27] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. Does object recognition work for everyone?. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2017–53.
[28] John Wiley & Sons. 2019. 115, 16 (2018), E3635–E3644.
[29] Ilan H Meyer. 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. Psychological bulletin 129, 5 (2003), 674.
[30] Ilan H Meyer. 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. Psychological bulletin 129, 5 (2003), 674.
