2024

Sexism, Racism, and Classism: Social Biases in Text-to-Image Generative AI in the Context of Power, Success, and Beauty

Eva Johanna Gengler
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, eva.gengler@fau.de

Recommended Citation

Gengler, Eva Johanna, "Sexism, Racism, and Classism: Social Biases in Text-to-Image Generative AI in the Context of Power, Success, and Beauty" (2024). *Wirtschaftsinformatik 2024 Proceedings*. 48.
https://aisel.aisnet.org/wi2024/48

Sexism, Racism, and Classism: Social Biases in Text-to-Image Generative AI in the Context of Power, Success, and Beauty

Research Paper

Eva Johanna Gengler[1, 2]

[1] FAU Erlangen-Nürnberg, Institute of Information Systems, Nuremberg, Germany
{eva.gengler}@fau.de
[2] enableYou Consulting GmbH, feminist AI, Neunkirchen am Brand, Germany

Abstract. This study examines the manifestations of sexism, racism, and classism in the output of six text-to-image generative AI systems within the constructs of power, success, and beauty. A total of 180 images were generated using three prompts for each AI tool. Our analysis focused on detecting gender, racial, and class biases, as well as age discrimination. The findings reveal an underrepresentation of women and People of Color across the generated images. Additionally, the tendency to depict women in a sexualized manner was prominent. Data also indicated a bias towards younger depictions of women relative to men and People of Color relative to white individuals. The images overwhelmingly represented individuals as belonging to a higher socioeconomic class, pointing towards a systemic bias within AI systems towards privilege.

Keywords: *Generative AI, text-to-image, sexism, racism, classism*

1 Introduction

The advent of large generative AI models such as ChatGPT, Midjourney, and Stable Diffusion marks a significant milestone in technological advancement, reshaping the landscape of digital communication, visualization, and creation with their generative capabilities (Hacker et al., 2023). Today, especially text generation and text-to-image generation AI models are widely used within organizations and by millions of individuals (Bianchi et al., 2023). Consequently, the generated content – shaped by societies and their user's world views – can in return shape user's perceptions (Newman & Schwarz, 2024) and if spread at scale, can have an impact on societal world views.

AI text generation builds on large language models, which are systems trained on string prediction tasks to forecast the likelihood of a token based on its context, operate unsupervised, and produce scores or string predictions upon receiving text inputs (Bender et al., 2021). They are frequently based upon pretrained representations of word distributions, known as word embeddings (Bender et al., 2021). Similarly, image generation builds on large corpuses of labeled und unlabeled images. To train these text and image generative systems, frequently, vast amounts of text and image data are being extracted from the internet (Tacheva & Ramasubramanian, 2023).

This data includes biases, stereotypes, and discrimination against marginalized people, frequently, including sexism, racism, and classism. As predominantly young male users (Barera, 2020; EJO, 2018; Mitchell, 2016) and those from high economic backgrounds produce data on social media, Wikipedia, and other outlets, these tend to oversample the views of privileged people (Hargittai, 2020). Recently, numerous instances of biased, stereotypical, and discriminatory AI generative imagery have appeared in media outlets (e.g., Thomas & Thomson, 2023; zdfheute, 2023), on social media platforms (Krawczyk, 2023), and in research (Bianchi et al., 2023; Hosseini, 2024; J. Zhang & Verma, 2021). For instance, whiteness has been reinforced as an ideal (Bianchi et al., 2023), racial and gender disparities have been amplified in images of occupation (Bianchi et al., 2023; Zhou et al., 2024), and stereotypical gender expressions and appearances (Zhou et al., 2024).

Generative AI's detrimental effects extend beyond biases, such as perpetuating neo-colonialism (Fischer, 2023) and the exploitation of labor, exemplified by the underpayment of Kenyan workers involved in the development of ChatGPT (Billy, 2023). The environmental cost of operating these AI models is considerable, demanding vast amounts of computation, electricity, and water (Bender et al., 2021; Fischer, 2023; Kenthapadi et al., 2023; Ludvigsen, 2022). The dual burden of escalating environmental and financial costs unjustly impacts marginalized communities, which are less likely to reap the benefits of generative AI and are more susceptible to the adverse environmental ramifications of the models' resource consumption (Bender et al., 2021). Thus, generative AI has become a driver for the privileged and a risk for those that are marginalized on various levels. However, the critical AI studies underscore that the systemic issues of algorithmically induced inequality and injustice have not only persisted but have been exacerbated by the latest generation of AI systems (Benjamins, 2021; Gordon, 2019; Raley & Rhee, 2023; Roberge & Castelle, 2021).

Research has extensively analyzed biases, stereotypes, and discrimination in AI generated text (e.g., Bender et al., 2021; Robinson, 2021; Smith & Williams, 2021), AI image generation, however, so far has merely been addressed by a small number of studies (Bianchi et al., 2023; Hosseini, 2024; Zhou et al., 2024) mostly focusing on occupational imagery. As images can influence our perception of the world, of the truth, and of ourselves (Newman & Schwarz, 2024), we need to gain a better understanding of the way biases persist or are amplified in image generative AI. To understand the nature of sexism, racisms, and classism in text-to-image AI, we conducted an empirical study using six popular AI models for image generation among three contexts. These are *power* (traditionally associated with male gender (Charafeddine et al., 2020) and white skin (Dukler & Liberman, 2022)), *success* (traditionally associated with male gender (Heilman et al., 2004; McColl-Kennedy & Dann, 2000)), and *beauty* (traditionally associated with female gender (Forbes et al., 2007; Travis et al., 2000) and white skin (Hall, 1996; Mady et al., 2023)).

Consequently, this research constitutes empirical findings, by asking:

RQ: *To what extent do generative text-to-image AI models exhibit gender, racial, and class biases in the context of power, success, and beauty?*

2 Theoretical Background

2.1 Biased, Stereotypical, and Discriminatory AI Output

Though, easing services and making grammatically perfect texts readily available, generative AI also presents several significant risks. Firstly, AI generated content might produce hallucinated – or in other words false or absurd – content. Secondly, it can be intently used to produce misinformation such as deepfakes or propaganda. Thirdly, it can by itself produce biased, stereotypical, and discriminatory content. In this paper we focus on the latter risk.

Biases are currently an integral part of generative AI systems. Large language models, for instance, have been shown to reinforces racist, sexist, ableist, extremist, or other harmful ideologies (Bender et al., 2021). When deployed in use cases ranging from the prescription of medical treatment (Robinson, 2021) to text (Alba, 2022) or image generation (Smith & Williams, 2021), these biases may strengthen discriminatory effects in societies. A typical example for gender bias and stereotyping is in occupation scenarios. This is true both for text (Smith & Williams, 2021) as well as image (Bianchi et al., 2023; Hosseini, 2024; Zhou et al., 2024) generative AI. Moreover, AI generated images often present racial biases (Bianchi et al., 2023; Zhou et al., 2024). Studies find that image generation amplifies gender and racial occupation disparities compared with labor force statistics and Google images (Bianchi et al., 2023; Zhou et al., 2024). Moreover, women were generally depicted as younger and more smiling than men (Zhou et al., 2024), and whiteness was reinforced as the norm and the ideal (Bianchi et al., 2023). People affected by discrimination through more than one attribute such as women of Color, who are affected by both sexism and racism (Crenshaw, 1989; Shaw, 2014), face even higher disadvantages through AI (Buolamwini, 2017). It appears, that this phenomenon is now being mirrored by contemporary generative AI systems (Tan & Celis, 2019; zdfheute, 2023).

Overall, generative AI has been shown to produce biased outcomes regarding ageism, sexism, racism, ableism, classism, conservatism, urbanism, as well as anachronism and perpetuate stereotypes in both text and image (Alba, 2022; Bianchi et al., 2023; Thomas & Thomson, 2023). When used, text-to-image models can both propagate unfair social representations and pose the risk of being used to aggressively market conservative ideologies (Vice et al., 2023).

2.2 Reasons for Discriminatory, Stereotypical, and Biased AI Outputs

The reason for biased, stereotypical, and discriminatory AI outcomes are complex: They firstly, reside in the societal system, secondly, in the people involved in training and making decisions about AI, and thirdly, in the training data itself. Fourthly, the internal functioning of generative AI play a role and fifthly, user prompting, which we will touch upon in the discussion. We will focus on reasons one to three in this study.

Firstly, Tacheva and Ramasubramanian (2023, p. 1) argue that “the dehumanizing and harmful features of the technology industry that have plagued it since its inception only seem to deepen and intensify. Far from a ‘glitch’ or unintentional error, these endemic issues are a function of the interlocking systems of oppression upon which AI is built." They argue the case for an AI empire, demonstrating that this interconnected and widespread global system is founded on heteropatriarchy, racial capitalism, white supremacy, and colonialism (Tacheva & Ramasubramanian, 2023). Moreover, it continues to extend its reach via the practices of extraction, automation, essentialism, monitoring, and control (Tacheva & Ramasubramanian, 2023), while being fueled by a largely Western understanding of technology (Aouragh & Chakravartty, 2016; Arora, 2016).

Secondly, in such a system the structures and the people creating these systems have an impact on its outcomes (Gengler et al., 2023). Heteronormativity often is a constant as exemplified by a culture of toxic masculinity in Silicon Valley (Chang, 2018; D'Ignazio & Klein, 2020; Shaw, 2014) leading towards systems that often function among others in a sexist and racist fashion (Noble, 2018). According to the World Economic Forum (2019), most developers in AI are men, women making up merely an estimated 26 % of workers in AI. Moreover, regarding the producers of data on platforms such as Wikipedia and Reddit, there is an overrepresentation of young male users from the Global North (Barera, 2020; Mitchell, 2016). Moreover, just 23 % of stories written in eleven European countries were written by women (EJO, 2018) and social media data is oversampled with views of privileged people (Hargittai, 2020). Moreover, due to the nature of social media, users tend to present themselves in a good light, and often publish only selected, edited, and unrealistic content (Tiggemann & Anderberg, 2020). Therefore, social media databases are never an unbiased representation of reality. These are all resources of data that generated AI is trained with. Thus, among others white supremacist and misogynist world views are overrepresented, exceeding their prevalence in the general population and laying the ground for misrepresentational training data (Bender et al., 2021).

Thirdly, the input data has a strong impact on AI outcomes (Gengler et al., 2023), as Birhane and Prabhu (2021, p. 1541) point out: "Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy." Especially as the data used for training is not representing our societies in an unbiased way. The data the models are trained on, shows an overrepresentation of hegemonic viewpoints and encoded biases from the internet (Bender et al., 2021). Especially the practice of word embeddings, vector-based language models determining semantic closeness of words, includes encoded social biases (Bartl et al., 2020; Bolukbasi et al., 2016; Tan & Celis, 2019). These biases include sexism (Bartl et al., 2020; Basta et al., 2019; Bolukbasi et al., 2016; Kurita et al., 2019; Sheng et al., 2019; Tan & Celis, 2019; H. Zhang et al., 2020; Zhao et al., 2019), racism (Kurita et al., 2019; Tan & Celis, 2019; H. Zhang et al., 2020), and ableism (Hutchinson et al., 2020). As large language models are often trained on word embeddings, these biases creep into the systems. Moreover, input data such as image descriptions can be biased and misleading (Bennett et al., 2021). We see similar problems, when regarding images. For instance, women are often sexualized in media portrayals (Vezich et al., 2017) and on social media (Slater & Tiggemann, 2016). However, not only biases but also misrepresentation is often a problem. As photos of men outweigh those of women across media by a factor of almost three (EJO, 2018; Rattan et al., 2019), there is simply a larger corpus of images depicting men, than women. This is also true for pictures depicting people from the Global North compared to those from the Global South, for instance. Generative AI is often trained on mostly uncurated media and texts, many of which reflect social biases (Jiang et al., 2023). Moreover, as the size of the datasets is growing, there is an accumulation of documentation dept (Bender et al., 2021). "Thus at each step, from initial participation in Internet fora, to continued presence there, to the collection and finally the filtering of training data, current practice privileges the hegemonic viewpoint. In accepting large amounts of web text as 'representative' of 'all' of humanity we risk perpetuating dominant viewpoints, increasing power imbalances, and further reifying inequality" (Bender et al., 2021, p. 614).

Finally, AI generated content has the tendency to "flatten out" controversial topics, voices from minorities, and diverse viewpoints towards a societally accepted or most strongly represented mean (Bender et al., 2021). This is particularly a problem for marginalized people.

3 Research Method

3.1 Process of Data Generation and Applied Text-to-Image Tools

We utilized six text-to-image AI generators, namely, DALL-E 2, Firefly, Leonardo.Ai, Magic Media, Midjourney, and Stable Diffusion, to generate images for three different contexts. We chose to use these platforms as they are among the most widely used AI image generators during the time of the study. Secondly, aiming for accessibility, each of them, except for Midjourney, can be accessed for free to a certain extent. Each of the generators is trained on a large corpus of images and accepts text-based prompts (in some cases possibly advanced with images as input and with different additional measures to impact the image generation such as styling) to generate images. While Midjourney is merely a subscription-based AI image generators, DALL-E 2, Firefly, Leonardo.Ai, Magic Media, and Stable Diffusion offer limited free access, too.

To generate our data, we used three consistent text prompts with no other input measures applied for all six models. Our three prompts consisted of "powerful people", "successful people", and "beautiful people". We aimed for three simple prompts that were nevertheless associated with specific societal, cultural, and gendered world views. *Power* is a concept often associated with masculinity. *Success* is frequently attributed to money and men, and *beauty* is often associated with a westernized view on female bodies. As we used DALL-E 2 via Chat GPT 4, which is a large language model, we enhanced the prompts to create an image with the words "create a picture of 'X'", "X" being the above-mentioned prompts. To build our analysis on some variety per prompt and model, we aimed for 10 images for each prompt and model. As DALL-E 2 creates only one picture at a time, we copied and pasted each prompt 9 times. Firelfy (an Adobe AI image generator), Magic Media (a generative AI model integrated in Canva Pro), and Midjourney each create four images simultaneously, so we regenerated each prompt three times, and subsequently merely used the first ten generated images. In Leonardo.Ai, we prompted eight images simultaneously, eight being the maximum number of simultaneous outputs at a time, and then two additional images. Finally, Stable Diffusion creates up to two images at a time without subscription. So, we regenerated the prompts four times each. After generation, we downloaded all images, numbered them from 001 to 180, and saved them in distinct folders according to their prompt and model. In total, we created 66 images for each prompt and 198 images in total. To reach a balanced data set, we selected the first ten images per tool, resulting with 180 images in our data set. We created all images in March 2024. A selection of one image per context is displayed in Figure 1.

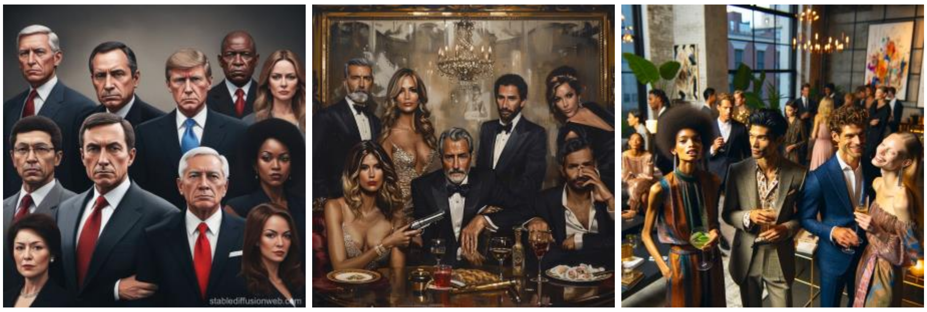

Figure 1. Selection of Three Exemplary Images (left to right) on the Contexts "Power" (Source: Stable Diffusion), "Success" (Source: Midjourney), and "Beauty" (Source: DALL-E 2)

3.2 Qualitative Analysis of the Generated Images

Researcher one analyzed the images integrating the perspective of a second researcher, if in doubt. Following Berger (2015), we share the positionality of researcher one. She is a white female researcher whose perspective is grounded in critical intersectional feminism. To analyze the content of the images, we created a table listing all prompts and tools with one row per image. The attributes assessed were perceived total number of people, number of both perceived women and men, number of both perceived white people and People of Color (PoC), the average perceived age of both perceived women and men, the average perceived age of both perceived PoC and white people, the appearance of both women and men, the appearance of both PoC and white people, the perceived class, and who was in focus. We analyzed the images as depicting PoC or white people based on the definition of Moses (2016) as encompassing all non-white groups and emphasizing the common experiences of systemic racism. The number of people was a numeric field. The age was divided in a range from 0 to 4 (0: 0 – 16, 1: 17 – 26, 2: 27 – 46, 3: 47 – 56, 4: 57 – 90). 2 and 4 were the largest ranges, as the assessing of the perceived age is most easy in very young or very old ages and thus, the ages in the middle are fuzzier and there were hardly any very old people depicted. The appearances, belonging to a certain class, and who was in focus were filled with short descriptions.

We then counted all perceived people on the images according to the attributes mentioned above. As some images showed many people in the background that were hardly recognizable, we concentrated our search on people in the front and merely those in the back that were not fuzzy. Moreover, we qualitatively assessed some more subjective factors: the perceived age of the people depicted, the appearance of the people especially in manner of clothing (e.g., business, elegant, casual, sexualized), and their class. In this process, the age was assessed as an average over the people depicted.

We then evaluated the results using several pivot tables concentrating on gender, racial, and age variations per prompt and tool as well as on the frequency of appearance description and class per prompt and tool.

4 Results

4.1 Systematic Gender, Racial, and Age Biases

Table 1. Gender Representation of our Dataset

Prompt	Tool	Total people	Total women*	Total men*	% women*
Powerful people	DALL-E 2	464	26	438	6%
	Firefly	20	9	11	45%
	Leonardo.Ai	89	16	73	18%
	Magic Media	423	36	387	9%
	Midjourney	28	3	25	11%
	Stable Diffusion	180	52	128	29%
Total		**1,204**	**142**	**1062**	**12%**
Successful people	DALL-E 2	218	67	151	31%
	Firefly	27	16	11	59%
	Leonardo.Ai	89	12	77	13%
	Magic Media	259	30	229	12%
	Midjourney	61	11	50	18%
	Stable Diffusion	107	18	89	17%
Total		**761**	**154**	**607**	**20%**
Beautiful people	DALL-E 2	146	87	59	60%
	Firefly	24	18	6	75%
	Leonardo.Ai	44	28	16	64%
	Magic Media	72	27	45	38%
	Midjourney	12	11	1	92%
	Stable Diffusion	36	26	10	72%
Total		**334**	**197**	**137**	**59%**
Overall total		**2,299**	**493**	**1,806**	**21%**

*perceived

Overall, our findings show an underrepresentation of women and PoC among all prompts and tools. Table 1 presents the gender distribution of individuals generated per prompt and AI tool. Across all categories, a total of 2,299 people were depicted and analyzed, with 493 (21%) being perceived as women and 1,806 (79%) perceived as men. For "powerful people," 1,204 people were produced with only 12% (N=142) being perceived as women. The highest proportion of women in this category was generated by Firefly at 45%. "Successful people" totaled 761 people, with women representing a higher percentage of 20% (N=154), where Firefly produced the highest percentage of women with 59%. The prompt "beautiful people" resulted in 334 people. It featured a significantly higher proportion of women at 59% (N=197), with Midjourney producing the highest percentage of women at 92%. These results indicate a gender bias in the image generation, with an overall tendency across all AI tools to depict men more frequently than women, except notably in the "beautiful people" prompt.

Table 2. Racial Representation of our Dataset

Prompt	**Tool**	**Total white* people**	**Total PoC***	**% PoC***
Powerful people	DALL-E 2	453	11	2%
	Firefly	13	7	35%
	Leonardo.Ai	39	50	56%
	Magic Media	263	160	38%
	Midjourney	20	8	29%
	Stable Diffusion	134	46	26%
Total		**922**	**282**	**23%**
Successful people	DALL-E 2	179	39	18%
	Firefly	24	3	11%
	Leonardo.Ai	80	9	10%
	Magic Media	243	16	6%
	Midjourney	54	7	11%
	Stable Diffusion	103	4	4%
Total		**683**	**78**	**10%**
Beautiful people	DALL-E 2	102	44	30%
	Firefly	14	10	42%
	Leonardo.Ai	19	25	57%
	Magic Media	50	22	31%
	Midjourney	8	4	33%
	Stable Diffusion	24	12	33%
Total		**217**	**117**	**35%**
Overall total		**1,822**	**477**	**21%**

*perceived

In our dataset assessing the racial diversity of images generated in response to the prompts "powerful people," "successful people," and "beautiful people," a notable variance in representation was observed as displayed in Table 2. Among the total number of people generated (N=922) by the prompt "powerful people", 23% were of individuals perceived as PoC. Leonardo.Ai produced the highest PoC representation at 56%. In contrast, for "successful people," the total number of generated people (N=683) had a significantly lower representation at 10%, with the highest representation by

DALL-E 2 at 18%. With "Beautiful people" generated people (N=217) showed greater racial diversity, with PoC representation at 35%, and the highest being by Leonardo.Ai at 57%. Overall, across all generated people (N=1,822), PoC representation was 21%, indicating a skew towards generating more images of white individuals across AI tools. These results underscore a disparity in racial representation, suggesting an area for improvement in the diversity of AI-generated imagery.

The results from our six text-to-image AI tools regarding perceived age in gender and race reveal implicit biases in generated images as depicted in Table 3. The average perceived age range for men depicted as "powerful people" spans from 57 – 90, whereas for women, it is 27 – 56 years, suggesting an association of power with age and gender disparity. Notably, for "successful people," both genders show a decreased average age range of 47 – 56 for men and 27 – 46 for women, indicating a younger demographic associated with success. For "beautiful people," the average age further declines to 27 – 46 years for men and 17 – 26 for women, illustrating a bias towards youth in standards of beauty. Racial analysis shows white individuals' age averaging higher (47 – 56) across all prompts, while PoC are on average depicted as younger (27 – 46), pointing towards a racial and age bias in the generation of these images. These data underscore the need for more inclusive and diverse representation in AI-generated imagery.

Table 3. Representation of Age in our Dataset per Gender and Race

Prompt	Tool	Avg. age men*	Avg. Age women*	Avg. age* white*	Avg. age* PoC*
Powerful people	DALL-E 2	57 - 90	47 - 56	57 - 90	47 - 56
	Firefly	47 - 56	27 - 46	27 - 46	27 - 46
	Leonardo.Ai	57 - 90	47 - 56	57 - 90	47 - 56
	Magic Media	57 - 90	47 - 56	57 - 90	57 - 90
	Midjourney	57 - 90	27 - 46	57 - 90	47 - 56
	Stable Diffusion	57 - 90	27 - 46	57 - 90	47 - 56
Avg. age powerful people		**57 - 90**	**27 - 56**	**57 - 90**	**47 - 56**
Successful people	DALL-E 2	47 - 56	27 - 46	47 - 56	47 - 56
	Firefly	27 - 46	27 - 46	27 - 46	27 - 46
	Leonardo.Ai	47 - 56	27 - 46	47 - 56	47 - 56
	Magic Media	47 - 56	27 - 46	47 - 56	27 - 46
	Midjourney	47 - 56	27 - 46	47 - 56	47 - 56
	Stable Diffusion	47 - 56	27 - 46	47 - 56	27 - 46
Avg. age successful people		**47 - 56**	**27 - 46**	**47 - 56**	**27 - 46**
Beautiful people	DALL-E 2	27 - 46	27 - 46	27 - 46	27 - 46
	Firefly	17 - 26	17 - 26	17 - 26	17 - 26
	Leonardo.Ai	47 - 56	17 - 26	27 - 46	17 - 26
	Magic Media	27 - 46	17 - 26	17 - 26	27 - 46
	Midjourney	17 - 26	17 - 26	17 - 26	17 - 26
	Stable Diffusion	27 - 46	17 - 26	17 - 26	17 - 26
Avg. age beautiful people		**27 - 46**	**17 - 26**	**17 - 26**	**17 - 26**
Total avg.		**47 - 56**	**27 - 46**	**47 - 56**	**27 - 46**

*perceived

4.2 Systematic Biases in Appearance and Class

The representation of gender and race varied notably across different prompts and AI tools. Women were typically shown in business outfits (N = 22), traditional clothing (N = 4), and elegant dresses (N = 4), and were sexualized (N = 3) in responses to the "powerful people" prompt, whereas men were predominantly portrayed in business attire (N = 44) or casual wear (N = 3), with no instances of sexualization. In the "successful people" category, women appeared in business (N = 26), elegant (N = 13), casual (N = 6) attire, and were often sexualized (N = 8). Men were again mainly shown in business outfits (N = 48). For "beautiful people," women were depicted in a mystical way (N = 15), casual (N = 15), elegant (N = 14), often adhering to a thin beauty standard, sometimes to the point of being very skinny (N = 7), and were frequently sexualized (N = 15). Men's images were largely casual (N = 12), fewer in business outfits (N = 4), and no instances of sexualization.

The dataset indicates that Firefly generated the most casually dressed women (N = 7) and no sexualized portrayals. Stable Diffusion produced the most sexualized representations of women (N = 9), followed by Midjourney (N = 8), and both Leonardo.Ai and DALL-E (N = 5). Notably, Firefly was the only tool that generated a single overweight individual, however, in the background, and one person with piercings. The analysis of facial expressions revealed a trend of unsmiling faces in "powerful people" images and smiling faces in "beautiful people" images. The data reflects entrenched biases in AI-generated images, where women and men are often depicted in gender-typical attire and roles, with women also portrayed in sexualized contexts. The almost complete absence of body diversity and alternative styles such as piercings suggests a narrow adherence to traditional beauty standards.

Regarding race, most white people and PoC were depicted in business clothing in both the "powerful people" (white N = 42, PoC N = 25) "successful people" (white N = 49, PoC N = 14) images. However, within the "beautiful people" images white people were most frequently depicted in a mystical way (N = 10) and PoC as casual (N = 9).

We also analyzed for class. Overall, predominantly people were depicted that appeared to be very privileged due to elegant clothing, make-up, jewelry, or other expensive items. DALL-E 2 and Midjourney merely created people that were perceived as privileged (N = 30). Whereas Firefly created also some people outside this category (N = 5). Likewise, Leonardo.Ai mostly created images with privileged people yet also some that depicted both people from lower classes/non-privileged groups and privileged people (N = 4), and in one case only supposedly people from lower classes/non-privileged groups (N = 1). Magic Media generated one picture of indigenous people, and some perceived as average (N = 2) or people from lower classes (N = 2). Finally, Stable Diffusion depicted merely privileged people in 24 cases images. This shows a bias towards privileged people which far outweighs the amount of privileged people in our societies.

5 Discussion, Limitations, and Path Forward

Our findings highlight the systematic presence of gender and racial as well as class-related biases across all utilized text-to-image generators and various prompts. In accordance, with less available images of women than men (Criado-Perez, 2020; EJO, 2018; Rattan et al., 2019) and in line with the findings of recent studies (Bianchi et al., 2023; Zhou et al., 2024), we see a strong misrepresentation of women across all AI tools and prompts to depict men more frequently than women, except notably in the "beautiful people" prompt. This highlights how women are scarcely associated with power, little more in success, but overrepresented when it comes to the concept of beauty. These findings mirror stereotypes predominantly existing in our societies and its media coverage today e.g., through the male gaze in movies (Mulvey, 2006). Moreover, our findings align with prior work (Bianchi et al., 2023; Zhou et al., 2024), underscoring the disparity in racial representation, generating far less images of PoC across all prompts and tools. Regarding the context of power, men of color were more often depicted as aggressive and with weapons than white men, even though white men were generated far more often in total. In comparison with a review by The New York Times (2020) on the faces of power in the United States being 80% white and only 20% PoC, the representation of PoC in our sample in the context of power with 23% seems to be almost accurate. This reflects what the sources of training data often is: Western and privileged countries. The lack of images of PoC is especially severe when it comes to the context of success, amplifying stereotypes against PoC. Regarding age distribution, women were overall depicted as younger than men, which is in line with Zhou et al.'s findings (2024) and is also true for images of women in news outlets and with roles in movies (Bazzini et al., 1997). This phenomenon was especially true for the context of beauty. Besides a young age, our data also presents "beautiful" women as very skinny, and often heavily made up. Images like these, if produced and spread at scale, might underscore the pressure on women to look young, thin, and according to unnatural beauty standards as is already underscored by social media (Slater & Tiggemann, 2016). Not only women were depicted as younger than men but also PoC as younger than white people, which highlights a bias against this community. Regarding the appearance of people generated, we see a bias toward depicting men – especially white men – in business clothing, and women – especially white women – either in business clothing or in elegant dresses and frequently sexualized. The sexualization of women in media portrayals is well documented and has had an impact on the pictures generated with contemporary systems (Vezich et al., 2017). The style of clothing is, moreover, associated with the appearance of coming from privileged classes, which is in accordance to the overrepresentation of privileged people producing images for social media (Hargittai, 2020).

Our study has implications for theory, extending the knowledge on the status of sexism, racism, and classism in contemporary text-to-image AI systems in the context of power, success, and beauty. We display how these systems mirror and amplify biases, stereotypes, and discrimination of our analog world. Future research should focus on this nexus by advancing AI systems to function in a more equitable and intersectional feminist way and by providing effective strategies for users to work against the biases ingrained in these systems building on extensive experiments in prompting. Moreover, education plays a crucial part in rising awareness about the non-objectivity of these systems and on how to critically reflect on their results.

Additionally, this research has implications for practice. Our findings underscore the importance of addressing the oppressive nature of generative AI systems in two ways: Firstly, the companies that develop, deploy, and use generative AI need to take steps towards less biased training data, set up guardrails for the functioning of their systems, and educate users about the risks of biased, stereotypical, and discriminatory content. Secondly, as these systems are in wide use today, users need to be empowered to have an impact on generative AI to create as few biased, stereotypical, and discriminatory output as possible. Future research should develop strategies and recommendations to advance the art of fair AI prompting towards more equitable outcomes.

Limitations of our work include the comparably small number of generated images, which might be a sample that is more biased than a larger dataset. We are positive that this is would not change our findings to a large extent in light of the many instances of biased AI images covered and shared on (social) media (e.g., Krawczyk, 2023; Thomas & Thomson, 2023; zdfheute, 2023) and in research (Bianchi et al., 2023; Hosseini, 2024; J. Zhang & Verma, 2021). Moreover, the selection of contexts and consequently, prompts might produce especially biased results. To mitigate this limitation, we prototyped several different short prompts mainly in DALL-E 2 that presented similar biased results, however, this process can still be hindered by a subjectivity bias. Likewise, subjectivity might have been introduced during the analysis of the data. As perceptions on gender, race, and age are subjective. We tried to minimize this bias by integrating the opinion of a second researcher, whenever the first researcher was in doubt. Additionally, we coded gender merely binary, though we are aware that this does not reflect the diversity of existing genders.

Moving ahead, future research ought to focus on how to prevent biased, stereotypical, and discriminatory outcomes of generative AI tools through alterations within the AI tools as well as prompting techniques. We want to educate and empower generative AI users. Thus, we have published a guide for fair AI prompting together with the Mittelstand-Digital Zentrum Zukunftskultur (Gengler et al., 2024). We aim at creating one of the many steps towards preventing harmful image generation at scale.

6 Conclusion

As generative AI continues to shape various facets of society and industry, the implications of its biases become increasingly significant. This study's exploration of generative AI's tendencies reveals critical biases: Women and PoC are notably underrepresented and often portrayed in ways that align with longstanding societal stereotypes. The younger depiction of these groups relative to their male or white counterparts, and the skewed representation of higher socioeconomic status, underscores a systemic bias in AI systems. These findings necessitate a rigorous evaluation of AI’s context and training datasets to ensure equitable and diverse representations and should prompt a discussion about the ethical development and deployment of such influential technologies.

References

Alba, D. (2022, December 8). OpenAI Chatbot Spits Out Biased Musings, Despite Guardrails. *Bloomberg.Com*. https://www.bloomberg.com/news/newsletters/2022-12-08/chatgpt-open-ai-s-chatbot-is-spitting-out-biased-sexist-results

Aouragh, M., & Chakravartty, P. (2016). Infrastructures of empire: Towards a critical geopolitics of media and information studies. *Media, Culture & Society*, *38*(4), 559–575. https://doi.org/10.1177/0163443716643007

Arora, P. (2016). Bottom of the Data Pyramid: Big Data and the Global South. *International Journal of Communication*, *10*(0), Article 0. https://ijoc.org/index.php/ijoc/article/view/4297

Barera, M. (2020). *Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia*. https://rc.library.uta.edu/uta-ir/handle/10106/29572

Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing* (pp. 1–16). Association for Computational Linguistics. https://aclanthology.org/2020.gebnlp-1.1

Basta, C., Costa-jussà, M. R., & Casas, N. (2019). Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 33–39). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3805

Bazzini, D. G., McIntosh, W. D., Smith, S. M., Cook, S., & Harris, C. (1997). The aging woman in popular film: Underrepresented, unattractive, unfriendly, and unintelligent. *Sex Roles*, *36*(7), 531–543. https://doi.org/10.1007/BF02766689

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Benjamins, R. (2021). A choices framework for the responsible use of AI. *AI and Ethics*, *1*(1), 49–53. https://doi.org/10.1007/s43681-020-00012-5

Bennett, C. L., Gleason, C., Scheuerman, M. K., Bigham, J. P., Guo, A., & To, A. (2021). "It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3411764.3445498

Berger, R. (2015). Now I see it, now I don't: Researcher's position and reflexivity in qualitative research. *Qualitative Research*, *15*(2), 219–234. https://doi.org/10.1177/1468794112468475

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *2023 ACM*

Conference on Fairness, Accountability, and Transparency, 1493–1504. https://doi.org/10.1145/3593013.3594095

Billy, P. (2023, January 18). *OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive | TIME*. Time. https://time.com/6247678/openai-chatgpt-kenya-workers/

Birhane, A., Prabhu, V., & Kahembwe, E. (2021). *Multimodal datasets: Misogyny, pornography, and malignant stereotypes*.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv:1607.06520 [Cs, Stat]*, 1–25. http://arxiv.org/abs/1607.06520

Buolamwini, J. A. (2017). *Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/114068

Chang, E. (2018). *Brotopia: Breaking up the boys' club of Silicon Valley*. Portfolio/Penguin.

Charafeddine, R., Zambrana, I. M., Triniol, B., Mercier, H., Clément, F., Kaufmann, L., Reboul, A., Pons, F., & Van der Henst, J.-B. (2020). How Preschoolers Associate Power with Gender in Male-Female Interactions: A Cross-Cultural Investigation. *Sex Roles*, *83*(7), 453–473. https://doi.org/10.1007/s11199-019-01116-x

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum, Vol. 1989*(1), 139–167.

Criado-Perez, C. (2020). *Invisible women: Exposing data bias in a world designed for men*. Vintage.

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. The MIT Press.

Dukler, N., & Liberman, Z. (2022). Children use race to infer who is "in charge." *Journal of Experimental Child Psychology*, *221*, 105447. https://doi.org/10.1016/j.jecp.2022.105447

EJO. (2018, May 15). Where Are The Women Journalists In Europe's Media? *European Journalism Observatory - EJO*. https://en.ejo.ch/research/where-are-all-the-women-journalists-in-europes-media

Fischer, J. E. (2023). Generative AI Considered Harmful. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–5. https://doi.org/10.1145/3571884.3603756

Forbes, G. B., Collinsworth, L. L., Jobe, R. L., Braun, K. D., & Wise, L. M. (2007). Sexism, Hostility toward Women, and Endorsement of Beauty Ideals and Practices: Are Beauty Ideals Associated with Oppressive Beliefs? *Sex Roles*, *56*(5), 265–273. https://doi.org/10.1007/s11199-006-9161-5

Gengler, E. J., Kraus, A., & Bodrožić-Brnić, K. (2024). *Faires KI-Prompting – Ein Leitfaden für Unternehmen*. BSP Business and Law School – Hochschule für Management und Recht. https://www.digitalzentrum-zukunftskultur.de/wp-content/uploads/2024/05/Faires-KI-Prompting-Ein-Leitfaden-fuer-Unternehmen.pdf

Gengler, E. J., Wedel, M., Wudel, A., & Laumer, S. (2023). Power Imbalances in Society and AI: On the Need to Expand the Feminist Approach. *Wirtschaftsinformatik 2023 Proceedings*. https://aisel.aisnet.org/wi2023/37

Gordon, F. (2019). Virginia Eubanks (2018) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: Picador, St Martin's Press. *Law, Technology and Humans*, 162–164. https://doi.org/10.5204/lthj.v1i0.1386

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123. https://doi.org/10.1145/3593013.3594067

Hall, K. F. (1996). Beauty and the Beast of Whiteness: Teaching Race and Gender. *Shakespeare Quarterly*, *47*(4), 461–475. https://doi.org/10.2307/2870958

Hargittai, E. (2020). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, *38*(1), 10–24. https://doi.org/10.1177/0894439318788322

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for Success: Reactions to Women Who Succeed at Male Gender-Typed Tasks. *Journal of Applied Psychology*, *89*(3), 416–427. https://doi.org/10.1037/0021-9010.89.3.416

Hosseini, D. D. (2024). *Generative AI: A problematic illustration of the intersections of racialized gender, race, ethnicity*. https://doi.org/10.31219/osf.io/987ra

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

Jiang, R., Kocielnik, R., Saravanan, A. P., Han, P., Alvarez, R. M., & Anandkumar, A. (2023, October 27). *Empowering Domain Experts to Detect Social Bias in Generative AI with User-Friendly Interfaces*. XAI in Action: Past, Present, and Future Applications. https://openreview.net/forum?id=GL7RDOru1k

Kenthapadi, K., Lakkaraju, H., & Rajani, N. (2023). Generative AI meets Responsible AI: Practical Challenges and Opportunities. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5805–5806. https://doi.org/10.1145/3580305.3599557

Krawczyk, L. (2023, October 1). *So fett-phobisch ist KI - Aus Selfies ein rpofessionelles Profilbild für LinkedIn erstellen lassen? Dank KI kein Problem. Oder doch?* LinkedIn. https://www.linkedin.com/posts/lisa-krawczyk-4501a3198_ki-kaesnstlicheintelligenz-fettphobie-activity-7100371178900316160-tJbp/?utm_source=share&utm_medium=member_desktop

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring Bias in Contextualized Word Representations. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 166–172). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3823

Lu, D., Huang, J., Seshagiri, A., Park, H., & Griggs, T. (2020, September 9). *Faces of Power: 80% Are White, Even as U.S. Becomes More Diverse—The New York Times*. The New York Times.

https://www.nytimes.com/interactive/2020/09/09/us/powerful-people-race-us.html

Ludvigsen, K. G. A. (2022, December 21). *The Carbon Footprint of ChatGPT: This article attempts to estimate the carbon footprint of the popular OepnAI chatbot called ChatGPT*. Medium. https://towardsdatascience.com/the-carbon-footprint-of-chatgpt-66932314627d

Mady, S., Biswas, D., Dadzie, C. A., Hill, R. P., & Paul, R. (2023). "A Whiter Shade of Pale": Whiteness, Female Beauty Standards, and Ethical Engagement Across Three Cultures. *Journal of International Marketing*, *31*(1), 69–89. https://doi.org/10.1177/1069031X221112642

McColl-Kennedy, J. R., & Dann, S. J. (2000). Success: What Do Women and Men Really Think It Means? *Asia Pacific Journal of Human Resources*, *38*(3), 29–45. https://doi.org/10.1177/103841110003800304

Mitchell, G. S., Jesse Holcomb and Amy. (2016, February 25). 1. Reddit news users more likely to be male, young and digital in their news preferences. *Pew Research Center's Journalism Project*. https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Moses, Y. (2016, December 7). Is the Term "People of Color" Acceptable in This Day and Age? *SAPIENS*. https://www.sapiens.org/language/people-of-color/

Mulvey, L. (2006). *Media and Cultural Studies: Keyworks* (M. G. Durham & D. M. Kellner, Eds.). John Wiley & Sons.

Newman, E. J., & Schwarz, N. (2024). Misinformed by images: How images influence perceptions of truth and what can be done about it. *Current Opinion in Psychology*, *56*, 101778. https://doi.org/10.1016/j.copsyc.2023.101778

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Raley, R., & Rhee, J. (2023). Critical AI: A Field in Formation. *American Literature*, *95*(2), 185–204. https://doi.org/10.1215/00029831-10575021

Rattan, A., Chilazi, S., Georgeac, O., & Bohnet, I. (2019, June 6). Tackling the Underrepresentation of Women in Media. *Harvard Business Review*. https://hbr.org/2019/06/tackling-the-underrepresentation-of-women-in-media

Roberge, J., & Castelle, M. (Eds.). (2021). *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*. Springer International Publishing. https://doi.org/10.1007/978-3-030-56286-1

Robinson, R. (2021). *Assessing gender bias in medical and scientific masked language models with StereoSet* (arXiv:2111.08088). arXiv. https://doi.org/10.48550/arXiv.2111.08088

Shaw, A. (2014). The Internet Is Full of Jerks, Because the World Is Full of Jerks: What Feminist Theory Teaches Us About the Internet. *Communication and Critical/Cultural Studies*, *11*(3), 273–277. https://doi.org/10.1080/14791420.2014.926245

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3407–3412).

Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1339

Slater, A., & Tiggemann, M. (2016). Little girls in a grown up world: Exposure to sexualized media, internalization of sexualization messages, and body image in 6–9 year-old girls. *Body Image*, *18*, 19–22. https://doi.org/10.1016/j.bodyim.2016.04.004

Smith, E. M., & Williams, A. (2021). *Hi, my name is Martha: Using names to measure and mitigate bias in generative dialogue models* (arXiv:2109.03300). arXiv. https://doi.org/10.48550/arXiv.2109.03300

Tacheva, J., & Ramasubramanian, S. (2023). AI Empire: Unraveling the interlocking systems of oppression in generative AI's global order. *Big Data & Society*, *10*(2), 20539517231219241. https://doi.org/10.1177/20539517231219241

Tan, Y. C., & Celis, L. E. (2019). *Assessing Social and Intersectional Biases in Contextualized Word Representations* (arXiv:1911.01485). arXiv. https://doi.org/10.48550/arXiv.1911.01485

Thomas, R. J., & Thomson, T. J. (2023, July 10). *Ageism, sexism, classism and more: 7 examples of bias in AI-generated images*. The Conversation. http://theconversation.com/ageism-sexism-classism-and-more-7-examples-of-bias-in-ai-generated-images-208748

Tiggemann, M., & Anderberg, I. (2020). Social media is not real: The effect of 'Instagram vs reality' images on women's social comparison and body image. *New Media & Society*, *22*(12), 2183–2199. https://doi.org/10.1177/1461444819888720

Travis, C. B., Meginnis, K. L., & Bardari, K. M. (2000). Beauty, sexuality, and identity: The social control of women. In *Sexuality, society, and feminism* (pp. 237–272). American Psychological Association. https://doi.org/10.1037/10345-010

Vezich, I. S., Gunter, B. C., & Lieberman, M. D. (2017). Women's responses to stereotypical media portrayals: An fMRI study of sexualized and domestic images of women. *Journal of Consumer Behaviour*, *16*(4), 322–331. https://doi.org/10.1002/cb.1635

Vice, J., Akhtar, N., Hartley, R., & Mian, A. (2023). *Quantifying Bias in Text-to-Image Generative Models* (arXiv:2312.13053). arXiv. https://doi.org/10.48550/arXiv.2312.13053

World Economic Forum. (2019). *Global Gender Gap Report 2020* (Insight Report, pp. 1–371). World Economic Forum. https://www3.weforum.org/docs/WEF_GGGR_2020.pdf

zdfheute. (2023). *Vorwürfe gegen virales KI-Tool für Headshots—KI sexualisieret Woman of Color*. https://www.instagram.com/p/CvM07apo5G3/?utm_source=ig_web_copy_link&igshid=MzRlODBiNWFlZA%3D%3D

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. (2020). Hurtful words: Quantifying biases in clinical contextual word embeddings. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 110–120. https://doi.org/10.1145/3368555.3384448

Zhang, J., & Verma, V. (2021). Discover Discriminatory Bias in High Accuracy Models Embedded in Machine Learning Algorithms. In H. Meng, T. Lei, M.

Li, K. Li, N. Xiong, & L. Wang (Eds.), *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery* (Vol. 88, pp. 1537–1545). Springer International Publishing. https://doi.org/10.1007/978-3-030-70665-4_166

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender Bias in Contextualized Word Embeddings. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 629–634). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1064

Zhou, M., Abhishek, V., Derdenger, T., Kim, J., & Srinivasan, K. (2024). *Bias in Generative AI* (arXiv:2403.02726). arXiv. https://doi.org/10.48550/arXiv.2403.02726