Towards a User-Centric Evaluation of Social Bias Mitigation in Text-to-Image Generative Systems

Master Thesis

Elise G.A. Mertens

2059488

February 2024

Internal Supervisor	**Second Internal Supervisor**	**External Supervisor**
Dr. ir. M.C. Willemsen	dr. Y.K. Leung	Dr. Fabian C. G. van den Berg
JADS	JADS	GroupM
M.C.Willemsen@tue.nl	Y.K.Leung@tilburguniversity.edu	fabian.van.den.berg@groupm.com

Jheronimus Academy of Data Science,
Data Science & Entrepreneurship

Towards a User-Centric Evaluation of Social Bias Mitigation in Text-to-Image Generative Systems

ELISE G.A. MERTENS, Jheronimus Academy of Data Science,
Data Science & Entrepreneurship
's-Hertogenbosch, the Netherlands

Although recent advances in generative artificial intelligence have notably improved image generation, their models often exhibit social biases, especially when generating images of people. Current attempts to debias such systems predominantly focus on auditing individual outputs rather than evaluating them within a broader system context aimed at specific usage-goals, neglecting the different perceptions of system aspects users might have. Through a user study, this research applies a user-centric evaluation framework to investigate how gender bias mitigation, under various settings, affects the perceived effectiveness of an image generation system in an ideation task within the advertising industry. The findings reveal that the perceived utility of such a system depends not only on the social diversity represented in the images but also on their quality, emphasising that diversity in itself also enhances perceived quality. While the mitigation strategy can improve perceived utility, its impact compared to unmitigated output is moderated by the inherent gender bias in the images and varies depending on the background of individuals, potentially influenced by their level of perceived unexpectedness of the output. This research underscores the complexities involved in designing effective systems in the context of social biases and the limitations of evaluating these systems in isolation. It moves a step closer to evaluating these systems in real-world scenarios, suggesting that future research should focus on the multifaceted nature of diversity and involve systems where users have greater control over generation and mitigation processes.

1 INTRODUCTION

Recent advances in text-to-image generation models, particularly diffusion models like Stable Diffusion [1] and DALL-E 3[2], have revolutionised the ease and capability of image generation, producing high-quality images from textual prompts in seconds [3]. This represents a significant advancement over previous methods, characterised by expanded capabilities and user accessibility [4]. However, due to the risks associated with their training and application, the increasing popularity of these models necessitates responsible design. The lack of diversity and inclusivity in the outputs of these models is a major source of concern. They frequently replicate and amplify harmful social stereotypes and biases, making the misrepresentation of gender, ethnicity, age, and other diversity groups a critical issue [5–7]. This issue is particularly severe in media production, such as advertising, where representation can have a major impact on viewer well-being and perpetuate stereotypes [8–10].

The difficulties in designing for fairness and inclusivity highlight the importance of using generative models responsibly. Due to the generative nature of these models, traditional approaches [11, 12] to fairness in data-driven decision-making are not entirely applicable. Because of the subjective nature of fairness [13], as well as the high variability in model [5] outputs and user perspectives, evaluating social biases in generative models is particularly difficult. This complexity is heightened by the variety of visual attributes in model outputs as well as the subjective interpretation of constructs such as gender and race [14, 15]. Furthermore, the massive scale of these models poses additional challenges. Earlier image generation models have relied on datasets limited in size and coverage compared to diffusion models [7, 16]. Removing bias through the pre-processing of multi-modal training datasets of enormous scale might be infeasible, and attempts such as filtering related to certain concepts continued to show biases on other

levels and led to a loss in generalisation ability [17, 18]. In addition, re-training or fine-tuning models with fairness constraints can be costly and can increase the carbon footprint of this field [19]

While various strategies for addressing social bias in generative models exist, research is still at an early stage. Studies often focus on occupational biases and aim to enhance social diversity in model outputs. They typically audit the models by classifying these outputs. However, these studies usually examine the system in isolation, not considering its real-life application or how it's perceived usefulness might change when applied. Additionally, they often overlook the underlying dynamics of mitigation methods, such as the impact of these strategies on the perceived quality of the outputs and the role of perceived diversity and surprise in ideation tasks. The practical effects and underlying dynamics of mitigation measures in real-world applications have not been thoroughly explored. This study seeks to bridge this gap by examining how employing social bias mitigation techniques affects the perceptions and utility of text-to-image generating systems in real-world scenarios. The central research question is: *How does the implementation of a social bias mitigation strategy impact the perceived effectiveness in a diversity-focused creative task for advertising?*

2 RELATED WORK

2.1 Diffusion Models for Image Generation

The general idea of generative modelling is that the approximate samples from a desired distribution are produced. A network is trained that models a distribution, like a distribution over images. The core idea is that a generative model's output comes from a distribution that is equal to that of the training data [20]. There are various generative models. One type of generative models are Generative Adversarial Networks (GANs). GANs initially revolutionised the field of image generation, and over time even allowed for the production of high-resolution image. The basic premise of the model is that it consists of a generator network that tries to produce samples that are as realistic as possible, and a discriminator network that aims to distinguish real samples from generated ones [21]. Another popular type of generative models for image generation are based on auto-regressive methods, like DALL-E [22] and CogView [23]. Central to this method is that these models view an image as a series of pixels and calculate its likelihood as the total of the conditional chances for each pixel. However, these methods are not without their drawbacks. GANs regularly suffer from complex training and issues like mode collapse, failing to reproduce a full extent of the data distribution [20]. Similarly, auto-regressive models have shown promise but are computationally intensive [3, 20, 24]. In contrast, diffusion models have recently emerged as a popular choice, offering simpler training procedures and the capability to generate a wide range of high-quality images [4].

A denoising diffusion probabilistic model, or Diffusion Model (DM) for short, can generate images through an iterative stochastic noise removal process. DMs are based on a denoising auto-encoder neural network and exist of two main processes: a forward and a backward diffusion process. During the forward diffusion process iteratively noise is added to training images, and the auto-encoder learns how noise alters the distribution underlying the images [25]. During the denoising process, reversing the inference it has learned, the trained neural network can construct an image, x , by subtracting the estimate of the random noise, $\tilde{\epsilon}$, from the actual noise z, as can be seen in Eq. 1. Furthermore, Because this is a complex problem, the noise removal happens iteratively in T steps, where the noise at each step can be defined as by Eq. 2 below.

$$x = z - \tilde{\epsilon} \tag{1}$$

$$z_{t+1} = z_t - \tilde{\epsilon}_t \tag{2}$$

To control the output of the T2I diffusion process, it can be conditioned on an encoded textual prompt. This involves classifier-free guidance [26], a technique used by multiple types of T2I models [27–29]. In this way, the estimate the denoising auto-encoder makes of the noise at a certain step, $\tilde{\epsilon}_t$, is pushed into the direction of the encoded prompt, c_p, at the extent of a guidance scale s_g, resulting in an image that resembles the prompt. Because the noise estimation depends on the model's parameters, θ, it can be indicated as

$$\tilde{\epsilon}_\theta(z_t, c_p) = \epsilon_\theta(z_t) + s_g(\epsilon_\theta(z_t, c_p) - \epsilon_\theta(z_t)) \tag{3}$$

In addition to the diffusion process, Stable Diffusion uses a pre-trained representational encoder, a CLIP model [30], to convert input text into embedding vectors. This is followed by the multi-step diffusion process using the neural network and a scheduling algorithm. To enhance efficiency, Stable Diffusion performs the diffusion process in a compressed latent space instead of directly on pixel images, using an auto-encoder for compression and reconstruction [31]. Because the diffusion and decoding process is influenced by the encoding of the prompt by a CLIP model and the model's training on various images, the social bias present in CLIP [30, 32] and the underlying dataset of Stable Diffusion, LAION-5b [18], can end up in the output of the model. As such, the CLIP model was found to frequently incorrectly classify male images to be related to crime, and images portraying black people as non-human, revealing significant biases in its image classification [32]. In addition, the LAION 5-b dataset consisting of 5.85 billion CLIP-filtered image-text pairs, has been found to contain content portraying malign stereotypes and racist and ethnic slurs [33].

2.2 Social Bias in image generation

In Text-to-Image (T2I) generative models, social biases often relate to the generation of images containing people. A frequently investigated topic of social biases is the depiction of gender and racial stereotypes for certain occupations, as they can readily be evaluated against existing workforce proportions [34, 35]. In T2I systems, research investigating prompts for different occupations is extensive, illustrating the perpetuation and amplification of various stereotypes [6, 7, 36–44]. Social bias that outputs might show are for example a prompt for “a photo of a face of firefighter" resulting in a majority of white people with masculine features, or a prompt for “a photo of a face of a housekeeper" giving darker skin tones and only feminine features [6]. Some more examples can be seen in Figure 1. This paper builds on this relatively large body of previous research on occupational social biases by connecting it to an application in the job advertising industry.

Identifying the source of bias in generative models and determining what part of the model to adjust is a complex task. Because this challenge exists in both language [45] and image models [46], it is difficult to understand the interaction of these biases in the latent space of multi-modal systems. Studies in image search [47] and captioning [48] have already highlighted these issues. More recent research by Friedrich et al. on bias in Stable Diffusion, the LAION-5B dataset [18, 33], and the CLIP model [30] identified biases in both the dataset and the model. The study found that biases in the dataset are often magnified in the outputs. However, pinpointing exact bias causes is difficult due to the intricate relationship between training data, objectives, and CLIP's inherent representations [41]. The task of identifying the source of bias becomes even more complex due to the variety of models and their versions. Although many models, such as Midjourney [49], CogView2 [7], and various versions of DALL-E [7, 37, 39, 43, 49], have been studied, the focus is often on Stable Diffusion. This is because strategies to mitigate bias often rely on the model's adaptability.

Fig. 1. Adapted illustration by the Bianchi et al. [6] how the prompt "A photo of the face of [occupation]" can amplify and perpetuate occupational biases, including gender and racial bias, for various occupations. These images were randomly sampled from 100 generated outputs by the authors.

Beyond just adjusting the prompt [36, 38, 44], Stable Diffusion is frequently analysed for social bias mitigation [6, 36, 40–42, 44, 50, 51]. Because it is open source, and has features like seed control for fair image comparisons make it a preferred model for this research as well.

Addressing social bias in text-to-image models is also complicated by the variety of relevant dimensions it encompasses. In recent research, the most frequently assessed features are gender [6, 7, 36, 37, 39–44, 49], skin colour [36, 37, 49], race [6, 40, 43, 44], and ethnicity [6, 39, 41]. Classifying such features also leads to difficulties and differences in evaluation methodologies. For example, for gender, only a few studies include non-binary options [6, 39, 49], and skin colour evaluations range from using the Monk Skin Tone Scale [37] to binary categories [36]. Assessments of race and ethnicity also vary, with some studies providing limited options [6], while others offer a broader range [44]. In addition, only few studies address the intersectionality of social biases, for example finding a overrepresentation of white males in occupational images [39], and an underrepresentation of "darker females" [49]. Lastly, whereas some focus exclusively on classification or comparisons through human evaluation [43, 49, 51], which can be costly, others use solely automated methods employing CLIP [39, 41, 42], which has its own inherent biases [30]. To overcome the limitations of both evaluation approaches, some researchers have adopted a hybrid method, examining the correlation between them. This mixed approach has revealed both similarities and discrepancies, leading to inconclusive results regarding the generalizability of these correlations [36, 37, 44].

2.3 Bias mitigation

Various approaches have been explored for mitigating social bias in the outputs of models. A common approach to do so is by augmenting the text prompt used. As such Bansal et al. [36] expanded prompts altering a prompt featuring an occupation, e.g."a photo of a lawyer", by adding "if all individuals can be a lawyer irrespective of their gender" to the prompt (see Figure 2). Although phrases like these seem to influence the image generation process, the authors do not advocate for the intervention as a solution to mitigate bias. Bianchi et al. [6] take a more explorative approach and look at the qualitative effects of carefully rewriting prompts. Besides experimenting with changing prompts to explicitly mention identity or demographic language, they also test the intervention of Bansal et al. [36]. They find that stereotypes persist regardless of these interventions, and highlight the complexity of mitigating such biases (an

example is shown in Fig 2). In line with this, Naik et al.[44] also suggests that using prompts that specifically mention gender, race, or age may not always be sufficient to mitigate biases in image generation. Furthermore, they discovered that image quality diminishes when prompts, explicitly mentioning a specific gender, do not align with the gender most represented in the baseline images generated from prompts without any gender specification.

Liu et al. [51] take a combined mitigation approach by using automated prompt engineering to add cultural context, and fine-tuning Stable Diffusion on a culturally curated dataset. They find that this leads to culture-related outputs that are less offensive and more culturally-accurate. Another paper that focuses on cultural biases by Struppek et al. [50] looks at how the insertion of non-Latin characters into prompts can introduce cultural stereotypes in generated images. The approach they take focuses on the text encoding and unlearning the representations of these non-Latin characters. Although both studies show promising results and attain there goal in acquiring desired outputs related to cultural representations, the generalizabitlity of these findings seems to be limited. Whereas the first approach relies on a specifically curated dataset for a selection of cultures, the latter does not go beyond mitigating biases strictly linked to a defined set of characters.

Three recent papers have shown promising attempts at mitigating social bias in text-to-image generative models. All argue that previous work has been more data-oriented, where the focus is on retraining or fine-tuning models on labelled datasets. However, creating a bias-free dataset is generally infeasible and hinders the model to sub-sequentially be used for a large variety of downstream tasks [17]. Orgat, Kawar, and Belinkov [42] demonstrate that updating the model's cross-attention layers by editing their projection matrices can change implicit assumptions that generalise to unseen classes. These layers connect visual information to textual information in the DM. They find that their method is successfully able to reduce occupational gender bias, while the overall quality of the images remains unaffected. However, the applications of the results of the paper seem limited. The authors stress that the learning rate influences the specificity and generalizability of the findings. For example, when using the model to mitigate stereotypes to be found on the job market, every different occupation seems to take in a different learning rate. The authors base this on the male female division according to labour statistics, but find that this might not be the best statistic to base the learning rate on as it does not reflect the occurrence of gender in the produced pictures. Furthermore, determining learning rates for biases that might not come with prior expected distributions, like race or age, can be complex.

Like Struppek et al. [50], Chuang et al. [40] use a method focused solely on the text encoder. They describe their method as a standardised preprocessing step to project out biased directions in a text embedding using a calibrated projection matrix. They apply their intervention to mitigating gender and race biases for occupations both for classification and image generation. They show their method is effective and after training also extend to unseen occupations, without having to calibrate a specific learning rate. Although the authors stress that their method is computationally efficient, they do not comment on how readily the preprocessing step can be applied and adjusted. They also do not investigate how their method affects the quality of images.

The most promising mitigation method out of the three by Friedrich et al. [41] allows the user to directly instruct the model on the direction of bias during deployment. Key to their "fair guidance" mitigation approach, is a method called Semantic Guidance (SEGA) [52]. This method is widely applicable, as it can be applied to any generative architecture using classifier-free guidance. The fair guidance extends the classifier-free guidance in the noise estimation as described in Eq. 3, to $\tilde{\epsilon_\theta}(z_t, c_p, c_e)$, as can be seen in Eq. 4. Here γ further relies on attribute expressions e_i and scaling factors s_{e_i} with a certain direction of guidance. The "Fair guidance" the authors propose, entails the further conditioning of the image generation beyond the initial prompt, through, in the binary case, steering, with s_{e_2} the image generation toward one concept, e_1, e.g. "female person", and steering, with s_{e_2}, away from another, e_2, e.g. "male person". To ensure

equal proportions of both concepts in the output (the authors' definition of fair), the direction of the steering randomly switched this direction with 50% chance. However, this scale can also be adjusted manually to set proportions as desired.

$$\tilde{\epsilon}_\theta(z_t, c_p, c_e) = \tilde{\epsilon}_\theta(z_t, c_p) + \gamma(z_t, c_e) \tag{4}$$

The result of the intervention is the editing of dedicated features in isolation, without requiring additional training, or external guidance. The paper introducing SEGA also highlights how the quality of images becomes better by the removal of uncanny artefacts, although this evaluation does not specifically focus on the mitigation of gender biases. Most importantly, evaluating their intervention by applying it to occupational gender bias, the authors are successfully able to shift the gender appearance away from existing biases over multiple occupations. An example of their use of the SEGA intervention compared to a baseline and two prompt engineering interventions can be seen in Fig. 2. The authors also provide further qualitative evaluations for applying the methodology to race, age, heteronormativity biases. Furthermore, the authors propose that their Fair Diffusion as part of a framework that can be deployed. The idea is that a user gives a prompt out of which a biased concept can be recognised. Depending on the biased concept, the fair instructions that have been related to the concept, so e_i, s_e, and the proportion desired, will be translated into fair guidance to the encoded text, resulting in debiased output.

Fig. 2. A comparison of different image editing techniques, adapted from Friedrich et al [41]. From left to right, the first image shows a baseline image of Stable Diffusion v1-5 for "A photo of the face of a fighter". The second image shows a simple intervention by changing "firefighter" to the gendered "female firefighter" in the prompt. The third image shows the intervention proposed by Bansal et al., extending the prompt with "if all individuals can be a firefighter irrespective of their gender". The last image shows the mitigation proposed by Brack et al. [52] that was used by Friedrich et al. [41]. It is evident how compared to the other interventions, the semantic guidance (increasing "female" and decreasing "male"), leaves the image composition the same while also successfully altering the gender appearance.

2.4 Inclusive Design

The previously discussed studies have sought to reduce bias in image generation models, primarily evaluating these models in isolation rather than as part of a larger system. Although Friedrich et al. (2023) [41] proposed how their solution works at deployment, they disregard user experience in their evaluations and use a model to classify the gender of the people the model output. The studies often overlook the fact that these models function within a system that includes an interface for interactive input and output control. Importantly, they tend to focus on evaluating outcomes retrospectively, rather than looking proactively how using such tools and interfaces can shape the creation process towards diversity and inclusivity.

Designing for such a co-creative user experience involving generative AI, is a novel field. A recent study by Weisz et al. [5] outlines design principles to achieve positive synergy in human-AI collaborative systems, emphasising that they are inherently designed for generative variability, producing multiple and diverse outputs for a given input. The authors advocate for a explorative and co-creative process, rather than aiming for the model to produce a perfect output. A lack of diversity in such a process might cause harm by displaying forms of social bias and stimulate unconscious stereotypes a user might hold.

A practical co-creative application in which bias in T2I provides opportunities and risks is in the advertising industry. Going beyond generating the exact image someone had in mind, T2I systems can prove both harmful and useful tools for generating ideas and inspiration for a advertisements. For example, a professional working in the advertising industry might want to get inspiration for an image for a job advertisement. In such a situation, a T2I model that produces biased output can be harmful, or be perceived as not useful. There is also the risk that the model's output may unknowingly reinforce the user's implicit biases, potentially leading to harmful effects in the final product. Ideally, a T2I model with effective bias mitigation would not only provide unbiased and helpful outputs for inspiration but also help users recognise and overcome unconscious stereotypes.

2.5 User-centric evaluation

In order to investigate how the mitigation of an image generation model can facilitate a user experience that is useful for an application like idea generation, a new evaluation approach is warranted. Current evaluation strategies classifying the outputs might give an indication of the displayed diversity in the model outputs, but do not account for the factors influencing the user experience and their perceptions of the system and output. Therefore, to evaluate the dynamics behind perceived usefulness of a T2I model, this study has drawn inspiration from evaluation techniques used in the field of recommender systems. Within the field, Knijnenburg et al. [53] have proposed a User-Centric Evaluation Framework that can be widely applied. The framework addresses the need to extend beyond the goal of accuracy, focusing on the subjective evaluation of interaction with a system. It recognises that aspects other than accuracy, including system features and personal or situational elements, can influence satisfaction. The framework suggests the use of evaluation metrics that account for attitudes towards these system aspects.

In the context of T2I generation, this could entail going beyond the purpose of generating the desired output as accurately as possible, and assessing the tool within different contexts. An example of this would be a co-creative tool that aims to inspire the user. Consequently, the system should prioritise factors enhancing the ideation process. These include variety, quality, and novelty [54]: useful ideas are different from each other, meet the specifications set, and are unexpected. The User-Centric Evaluation Framework can take these factors into account, including the dynamics between them. For example, while model output can be more diverse, this might not always be useful for ideation if this is at the cost of the image quality or if the result is entirely in line with the user's expectations. Moreover, rather than solely auditing the output, this approach takes into account users' varying attitudes towards system aspects, such as perceived system utility, output quality, and level of surprise [55].

It is essential to look at the influence of perceived social diversity in order to investigate the usefulness of social bias mitigation. The effect of social diversity on this perceived usefulness has been investigated in information retrieval research. For example, image search results that are diverse on account of gender and race have been found to result in significantly higher ratings when it comes to overall satisfaction using the tool and the likelihood of using it again in the future [56]. However, the contrary has been also been found. For instance, people were found to generally judge the relevance of the stereotype-confirming search results higher than documents that disconfirm them [57]. Moreover,

increasing social diversity can disguise or perpetuate existing biases. For example, using a display of diversity through a mosaic of individual photographs showing racially different individuals has been criticised for being “insufficient for meaningful social change" [58]. In the field of recommender systems, social diversity is not an active topic. However, more general diversity of recommendations is, and has been found to have multiple functions. This includes the prevention of overfitting, enhancing user experience, and boosting satisfaction [59, 60]. As such, it was found that balancing individual recommendation quality with recommendation-set diversity can improve their subjective evaluation [61]. This balance, however, presents a trade-off between utility and diversity, as excessive diversity can lead to reduced perceived utility [62].

A suggested mechanism on how diversity influences perceived utility, is through improving the perceived recommendation quality, which in turn is linked to increased perceived system effectiveness [63]. The aforementioned mitigation studies have also revealed a complex relationship between increased diversity and quality. Some findings suggest that efforts to reduce bias, such as expanding prompts, may decrease the measured resemblance to training data [44], and may additionally reduce the manually evaluated alignment between the prompt and output [51]. However, other mitigation methods using automatic evaluation find that the quality remains relatively unaffected [42, 50, 52]. Still, studies often do not particularly assess the quality in the specific context of social bias mitigation [41, 42], and rely solely on automatic evaluation, which has been shown to not directly align with the with human perception of quality [64].

Another way in which increased social diversity in T2I systems can aid in ideation, is through increasing perceived novelty [54]. Generally, recommender system research has found that high levels of novelty have been found to adversely affect user satisfaction and preferences [59]. Similarly, if a T2I system produces outputs that are excessively novel or unusual, they might not always be useful. Conversely, however, novelty has also been found to have a positive impact on user satisfaction [65]. Moreover, it has been proposed that increased diversity in recommendations leads to unexpectedness, which is linked to serendipity or pleasant surprise [66]. What role surprise can play in idea generation through T2I systems is therefore unclear and remains an under-explored area of research.

2.6 Research questions

Given all these factors that could influence perceived utility of a T2I system for ideation, and given that adjusted diversity might relate to all these factors, it is interesting to see how the mitigation aimed at increasing diversity influences these dynamics. This research will investigate the promising debiasing strategy proposed by Friedrich et al. [41], utilising the User-Centric Evaluation Framework to assess their equal-gender mitigation approach. In addition, this study will further expand on their research in two significant ways.

Firstly, a novel mitigation setting will be introduced, focusing on the overrepresentation of underrepresented gender groups, rather than striving for equal representation. Inspired by research into occupational biases in information retrieval [34], this approach aims to explore the effects of varying representation levels on user perceptions, particularly examining the impact of overrepresentation in contrast to underrepresentation and equality. This strategy not only tests the effectiveness of the “equal proportions” fairness assumption proposed by Friedrich et al. [41] in an ideation context, but also allows for a deeper examination of how the dynamics of such mitigation may influence system aspects like perceived quality and surprise. Secondly, the research will assess the influence on perceived utility of gender mitigation when incorporating both gender and racial diversity in the baseline. This expanded approach not only aims to further validate their findings, but also to explore the intersectionality of these dimensions of social diversity [6, 39, 49]. Furthermore, by examining how users from different backgrounds perceive the results, the study seeks to

understand the varying impacts social bias mitigation might have on users with different racial and gender identities, as this can too has been found to influence their perception of a system [34, 56].

Taken together, the current research aims to answer the three following research questions:

- **RQ1**: How do users' perceptions of diversity, quality, and surprise influence the perceived utility of the T2I system?
- **RQ2**: How does a social bias mitigation strategy influence the perceived utility of the T2I system?
- **RQ3**: How does the social bias context of the unmitigated images, as well as the personal context of the user, influence the perceived utility of the T2I system?

3 METHODOLOGY

3.1 Methodology

In this study, the mitigation approach investigated will be Fair Diffusion as proposed by Friedrich et al. [41], using SEGA [52] to mitigate gender occupational bias. To keep the setup as similar as possible to theirs, the T2I model that will be used is Stable Diffusion (SD) v1.5 framework [29]. Following their experimental approach, first a set of images were generated without intervention. The prompt "A photo of the face of a [occupation]" was used to generate 3 batches of 9 images for each occupation. Only images that were detected to contain human faces [67] and passed the Not Safe For Work filter SD offers [68], were stored together with their seed and model settings.

3.2 Occupational bias subjects and dimensions

The chosen occupations for this study were determined by the gender occupation bias identified in SD, as detailed by Friedrich et al. [41], and occupational demographics from the U.S. Bureau of Labor and Statistics (BLS) [69]. This selection aimed to mirror the gender biases in LAION 5-B, a western-centric dataset [70], as closely as possible to actual demographic distributions [18]. The biases were categorised into three bias groups: female, male, and both. While the primary focus is on gender as a dimension of social bias, the study expands its scope to include racial aspects, following a methodology similar to that of Metaxa et al. [34]. To explore the impact of race, occupations with specific racial percentages were identified using BLS data [69]. Given that white people are the most represented group in the underlying dataset, a straightforward binary categorisation was chosen, differentiating between white individuals and People of Colour (POC), to also maintain a manageable complexity in the experimental design. The BLS data has been aligned as much as possible with both persistent gender, as well racial stereotypes in the model's output according to varying studies [39, 41, 44]. Lastly, a factor that has been taken into account is if the images are mostly in colour, and clearly display faces. For this reason the occupation 'photographer' was not selected, as many faces were obstructed by cameras. The final selection and relevant statistics can be found in Figure 1.

3.3 Mitigation approaches

Using the same prompt, model settings, and seed as in the *baseline condition*, Fair Guidance was used during the image generation process steering in directions -"male person" and + "female person" respectively, mimicking the setup proposed by Friedrich et al [41] that they found to be most effective for mitigation. The *equal mitigation condition*, will also follow the fairness assumptions suggested by Friedrich et al. [41] that an "equal proportion of female- and male-appearing images is desired". Based on the their sampling method, images within each batch will be steered in

Occupation	Bias groups	Gender statistics		Race statistics
		SD v1-5 [41] (man - woman)	U.S. Census [69] (% total employed women, average=46.8, std = 27.7)	U.S. Census [69] (% total employed white, average=77.0, std = 9.5)
Nursing Assistant	Female - POC	1 - 250	90.0	55.4
Dental Hygienist	Female - White	0 - 250	96.3	91.4
Security Guard	Male - POC	243 - 7	24.3	54.9
Electrician	Male - White	250 - 0	2.2	92.8
Cook	Both - POC	154 - 96	38.4	69.4
Producer	Both - White	100 - 150	44.2	82.9

Table 1. Six occupations were selected covering the 3 different gender biases, and the 2 different racial biases. For each occupation, the amount of baseline images generated by "A photo of the face of a [occupation] classified as "man" or "female" in the study by Friedrich et al. [41] have been displayed. In addition, the proportions of people that are women and people of colour according to the BLS [69] demographics have been given, accompanied by the average and standard deviations of these statistics.

either direction with 50% chance. Because of the identical prompt, settings, and seed, the output stay as close as possible to the baseline in all other dimensions [52].

To explore the dynamics of the mitigation method, another strategy, *inverse mitigation*, will be implemented as a control condition. Inspired by work of Metaxa et al. [34], this approach aims to counteract the gender occupational bias in the output by steering it in the opposite direction. The rationale for choosing this as a control condition is twofold: it might prove beneficial for ideation processes, and it allows for an further examination of the potential impacts the mitigation might have on systems aspects such as perceived quality and surprise. In cases where occupational gender bias is both, this method will deliberately direct the results towards one of the gender categories (i.e. cook will be steered to female bias, and producer to male bias). For each batch, 8 of the images will be steered in the opposite direction, and 1 will be steered in the direction of the baseline bias.

3.4 Experiment design

A online user experiment was conducted as frequent in recommender systems research [53]. A sequential within-subject design was chosen to take into account different perceptions users might have, which is especially relevant regarding social biases. Participants will be subjected to the baseline and manipulations in random order, each for applied to different occupations. Care was taken to assure that the experiment was not too repetitive, and order effects were controlled for.

Rather than viewing the images separately or without the prompt, a 'screenshot' of a system was shown. This non-existent system, which can also be seen in Figure 3, displays nine images simultaneously. Per occupation, the 9 images are selected from one of three baseline batches to minimise interference and better replicate real-life conditions. The mitigated images for these batches have only been manually checked to ensure they are free from disturbing outputs. This setup is chosen to mimic existing interfaces, keeping the system and outputs as close as possible to real-life scenarios and to existing systems like Midjourney [71] and DALLE-3 [72], which are commonly used and often have limited control options in their interfaces. Additionally, the display of 9 images at the time facilitates the observation of clear diversity between images, and makes up for the setup not allowing for regeneration of images. The setup was also chosen to fit the current online survey setup as the static images make direct comparison between baseline and mitigation images possible by keeping the seed the same.

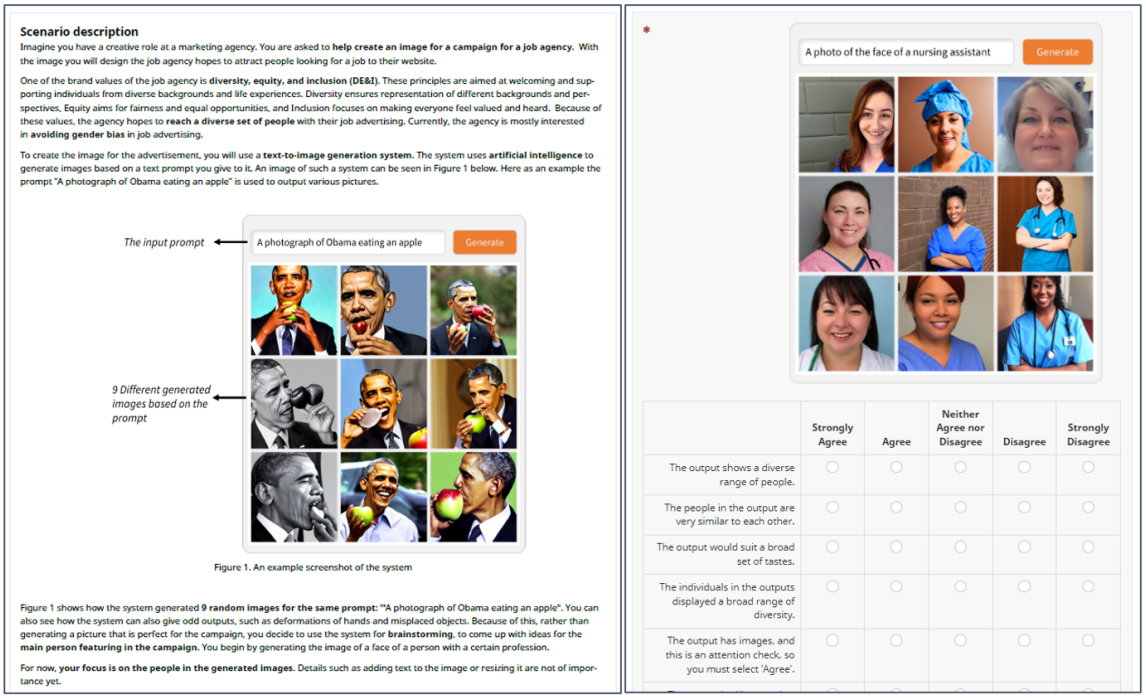

Fig. 3. Screenshots of the user experiment. Left: The scenario description participants viewed. This was also accompanied by a task description, which can be seen in Appendix A. Right: A page showing the system screenshot and the top of the questions (an overview of all questions can be found in Table 3. The participant would view three of such system screenshots and fill in the survey for each.

3.5 Participants

Participants were recruited via Prolific, which directed them to the survey platform LimeSurvey. Participants were paid £1.2 (on average the study took about 8 minutes) and were screened to target participants who were based in the USA and have specified their gender and ethnicity. In addition, Prolific allowed to select a balanced sample based on sex, to evenly distribute the study to male and female participants. The study was approved by the Ethical Review Board of the Eindhoven University of Technology. In total, 150 valid responses were received, passing at least 2 of the 3 attention checks. Still, 4 participants were excluded (the 2.5th percentile) that completed the study in less than 3.5 minutes, which was deemed too little too fully grasp the instructions and survey content. The mean age was 36.8 (SD = 13.3, min = 19, max = 77), with 72 females and 74 males, of which 8 participants identified as non-binary or trans. In addition, the distribution of the simplified ethnicity was as follows: 26 Asian, 16 Black, 85 White, and 21 mixed and other.

At the start of the experiment, the participants agreed to the informed consent which would take them to a description of the scenario, an example, and the task. The scenario outlined that the user has to imagine they are in a creative role at a marketing agency, tasked with creating an image for a job agency campaign, focusing on attracting job seekers while emphasising the agency's commitment to diversity, equity, and inclusion (DE&I). They will use a text-to-image AI system to brainstorm and generate images of diverse professionals, prioritising the portrayal of different backgrounds and avoiding gender bias, with less emphasis on additional image details at this stage. An example, unrelated to the actual task, was given to help users get familiar to the look and features of the T2I system, and to illustrate frequent

imperfections generated such as misplaced objects or deformed hands. The scenario and task description can be found in Appendix A. After this, participants had to complete two comprehension checks to see if participants had understood critical information relating to the goal of the job agency to reach a diverse set of people, and the task of brainstorming.

3.5.1 Study conditions. Each participant was subjected to the 3 conditions: the *baseline*, *equal mitigation*, and *inverse mitigation* settings in random order. For each mitigation, a different occupation was shown in the prompt and system output. All participants were on separate pages subjected to 3 occupations that together all three gender occupation bias conditions (Male, Female, Both) and both racial bias conditions (POC, White). As such, a participant could not have seen two occupations with similar gender stereotypes (e.g. both the security guard and electrician setting), or three occupations with the same racial stereotype (e.g. Dental Hygienist, Electrician, and Producer). An example can be seen in Figure 4. For all possible conditions see Table 2. In this way it was assured that every participant was subjected to every condition and social bias dimension.

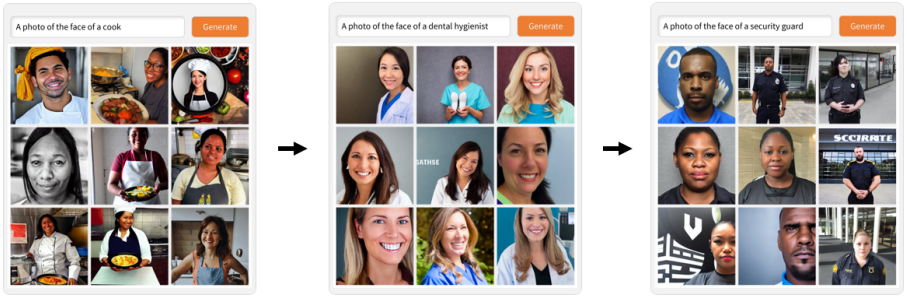

Fig. 4. An example of the system screenshots a participant evaluated. In line with Table 2, the respective participant would have been part of Group 1, the WPP racial bias group. Note that the order of what mitigation condition participants viewed was randomised too.

3.6 Survey design

After the scenario and task description, the participants were subjected to the 3 conditions. Each condition was separately accompanied by the same survey containing questions about perceived diversity, perceived quality, perceived surprise, and perceived utility of the system for the task. An overview of the questions can be found in Table 3. For all these questions, a 5-point scale ranging from "strongly disagree" to "strongly agree" was used. The participants were able to review the instructions during the trials. In addition, for every condition, one of the questions was an attention check.

The questions were adapted from various existing recommender system studies. For example, existing diversity questions [59, 63] have been adapted to fit social diversity, and caution has been taken to not use the term diversity in isolation. For example, the question "The recommendations contained a lot of variety" [63], was adapted to question D1 (Table 3). Similarly, questions for perceived utility [63, 65] have been adapted to focus more on the specific goal of this system (e.g. see question U3 in Table 3). The questions for perceived surprise have been adapted from perceived novelty and serendipity items [59, 65, 66], such that take into account that all generated images are to some extent novel. Lastly, the questions for quality [63] have been adapted to reflect both the perceived realness as well as the text-image alignment, to align with quality metrics used in prior research.

4 RESULTS

The user experiment was run in November 2023, one trial run for 20 participants, and one for 130. After the trial run only minor changes were made. Participants were randomly assigned to one of 6 groups, with each a unique combination of the baseline and mitigation conditions of the different gender occupational biases. Within each group, participants were again randomly assigned over the racial occupation bias conditions, another 6 groups. Because this resulted in a total of 36 groups, not every combination occurred in the study (See Table 2, Group 4, racial bias group PPW) However, this does not matter as every participant is individually exposed to all conditions and all biases. The order of presentation of the conditions was randomised, and did not result in any significant differences. An overview can be seen in Table 2, including the number of participants assigned to each group. The average time taken was 10.61 minutes (SD = 6.9).

Group	Conditions and Gender bias		Occupations	Racial Bias (in order of conditions)
1 (19)	Baseline	Female	Na (11), Dh (8)	PPW (4), PWW (3), PWP (4), WPP (3), WPW (1), WWP (4)
	Eq. Mitigation	Male	Sg (8), El (11)	
	Inv. Mitigation	Both	Co (11), Pr (8)	
2 (22)	Baseline	Female	Na (13), Dh (9)	PWW (4), PWP (1), PPW (8), WWP (2), WPP (6), WPW (1)
	Eq. Mitigation	Both	Co (15), Pr (7)	
	Inv. Mitigation	Male	Sg (9), El (13)	
3 (29)	Baseline	Male	Sg (13), El (16)	PPW (2), PWW (3), PWP (7), WPP (3), WPW (8), WWP (5)
	Eq. Mitigation	Female	Na (13), Dh (16)	
	Inv. Mitigation	Both	Co (15), Pr (14)	
4 (16)	Baseline	Male	Sg (8), El (8)	PWW (4), PWP (4), PPW (0), WWP (3), WPP (3), WPW (2)
	Eq. Mitigation	Both	Co (5), Pr (11)	
	Inv. Mitigation	Female	Na (10), Dh (6)	
5 (31)	Baseline	Both	Co (18), Pr (13)	PPW (5), PWP (7), PWW (6), WPW (2), WPP (7), WWP (4)
	Eq. Mitigation	Female	Na (14), Dh (17)	
	Inv. Mitigation	Male	Sg (18), El (13)	
6 (29)	Baseline	Both	Co (12), Pr (17)	PWP (6), PWW (1), PPW (5), WWP (6), WPP (5), WPW (6)
	Eq. Mitigation	Male	Sg (16), El (13)	
	Inv. Mitigation	Female	Na (17), Dh (12)	

Table 2. Display of all the groups and subgroups. Participants were randomly assigned to groups and viewed conditions in random order. Note the abbreviations for the occupations Na, Dh, Sg, El, Co, and Pr, respectively refer to nursing assistant, dental hygienist, security guard, electrician, cook, and producer. The numbers behind the groups and racial bias subgroups indicate how many participants were randomly assigned to these conditions. The numbers behind the occupations indicate how often the condition applied to this occupation was shown. For Racial Bias, P indicates POC, and W indicates White. An example of a group a participant could belong would Group 1 + Racial Bias Subgroup WPP. This means the participant saw the baseline images of the dental hygienist, the equal mitigation images of the security guard, and the inverse mitigation condition of the cook. An example of the images can be seen in Figure 4

4.1 Response Model and average effects

In order to make inferences about the perceived utility of the T2I system in relationship to other objective and subjective system aspects, confirmatory factor analysis (CFA) and structural equation modeling (SEM) have been applied using Lavaan[73] for R[74] and Mplus [75]. This is according to the guidelines for User-Centric Evaluation Framework [53].

CFA is used to test for construct validity of the used survey. It looks at how for each scale, e.g. diversity, every question relates to the latent factor that represents the scale. In order to perform the CFA, all survey responses were

Subjective aspects with items and codes		Full CFA Coef		SEM
		Coef.	R^2	Coef.
	Diversity (AVE = 0.753 , Alpha = 0.907)			
D01	The output shows a diverse range of people.	0.913	0.83	0.910
D02	The people in the output are very similar to each other.	0.695	0.48	0.697
D03	The output would suit a broad set of tastes.	0.887	0.79	0.885
D04	The individuals in the outputs displayed a broad range of diversity.	0.951	0.91	0.955
	Surprise (AVE = N.A., Alpha = N.A.)			
S01	The output had images that pleasantly surprised me.	1.112	NA	
S02	The output had images that were unexpected.	0.344	0.12	
S03	The system gives obvious suggestions.	0.198	0.04	
S04	The output had images I would not have thought to consider.	0.484	0.23	
	Quality (AVE = 0.782, Alpha = 0.885)			
Q01	The people in the images look realistic.	0.846	0.72	0.852
Q02	The output corresponds well with the prompt that was given.	0.654	0.43	
Q03	The people in the output have many features that look fake.	0.879	0.77	0.913
Q04	The system output contains multiple bad images.	0.874	0.76	0.887
	Utility (AVE = 0.693, Alpha = 0.839)			
U01	The system does not have a real benefit to me.	0.711	0.48	
U02	The system makes me more aware of my choice of options for the campaign design.	0.690	0.51	0.717
U03	The images in the output helped me to discover ideas to attract a diverse set of people.	0.888	0.79	0.915
U04	The system can help me in completing the task.	0.850	0.72	0.853

Table 3. Overview of subjective system aspects, their respective survey questions and codes. Per question, the initial CFA coefficients and R^2 have been shown. The SEM factor coefficients are significant with p <0.001. AVE is the average variance extracted, Alpha is Cronbach's alpha. All questions were answered with a 5-point Likert scale, ranging from strongly disagree to strongly agree. Questions D02, S03, Q03, Q04, and U01 were reverse coded.

converted to a 0-4 scale. The CFA treated the scores for each item as ordinal, and was able to in part confirm the proposed factor structure. To ensure convergent validity, some questions were excluded in a step-wise manner by investigating the Average Variance Extracted (The R^2 average for all items on a factor is > 0.5), making sure at least 3 items remained per factor. To ensure discriminant validity, high cross-loadings were taken into consideration. This iterative process resulted in dropping two survey items loading on quality and utility with low explanatory power, for example "The output corresponds well with the prompt that was given" ($R^2 = 0.42$) . It also resulted in dropping the surprise factor, as multiple questions had high cross-loadings and the Average Variance Extracted (AVE) was low. One of the items of surprise, however, had a relatively high individual R^2 and low cross-loadings and has been included in further analysis as an indicator variable.

The SEM model is shown in Figure 5. It can be seen as form of a CFA, in which the factors have been regressed on each other as well as the conditions of the experiment. Compared to other statistical models, the SEM has the unique ability to estimate both the measured factors and all hypothesised paths in a single model [53], allowing to make inferences about the causal structure of a model. Contextual bias factors of the images, i.e. gender and racial bias, have been related to perceived diversity and quality to see how these influenced the effects of the conditions. For interpretation, the baseline taken for this model is the equal mitigation condition with the both gender bias and the white racial bias. Like with the CFA, insignificant effects (p>0.05) were iteratively dropped from the model. For example, compared to the white racial bias, there was no significant interaction effect between POC racial bias and the effects of different conditions. Similarly, no interaction effects were found between the participant's racial and gender identity

and the biases in the output. However, for better interpretation, some insignificant effects have remained in the model as can be seen by the dotted lines in Figure 5. The fit overall fit of the model was good ($\chi^2(122) = 170.421$, p =0.003, CFI = .994, TLI =.998, RMSEA = .030, 90% CI : [0.018, 0.040]).

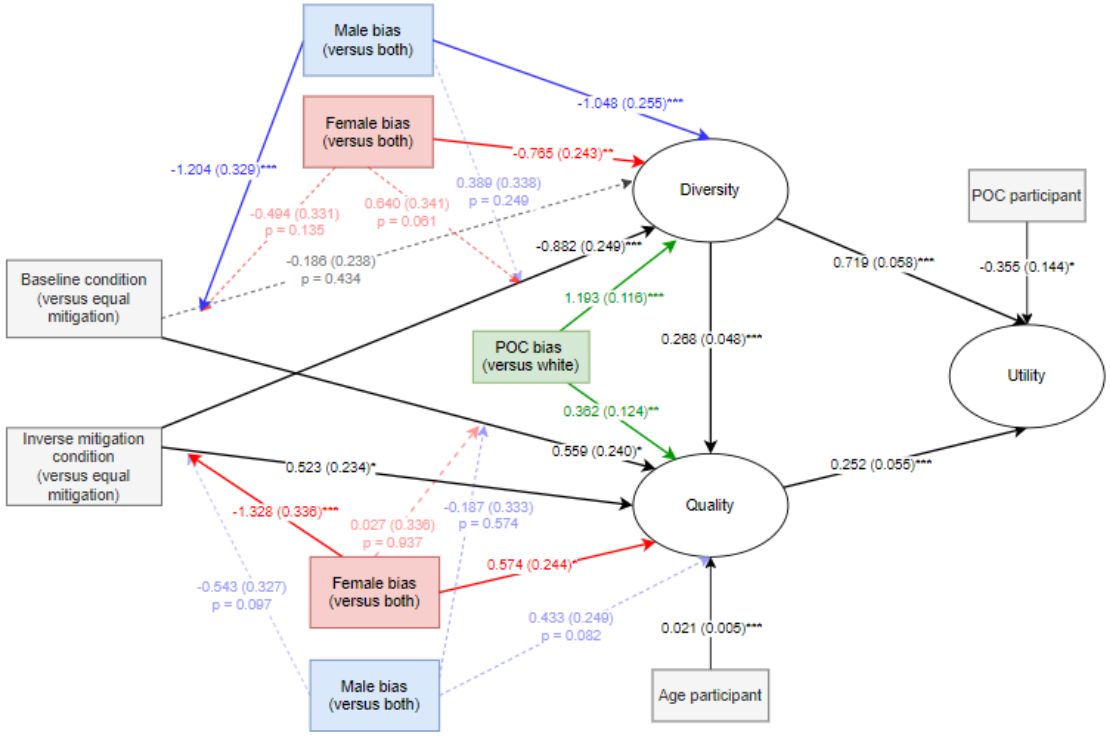

Fig. 5. SEM model showing the relations between the conditions and perceived system aspects. Arrows have their coefficients with standard error between brackets and p-values. If no p value is indicated, *** = p<0.001, ** = p<0.01, *= p<0.05. Dotted lines indicate relations with p>0.05. Here POC Participant indicates the effect of the indicated non-white ethnicity of the participant in the model.

Looking at the SEM model in Figure 5, it can be seen how be seen how through the perceived system aspects diversity and quality, perceived utility of the system increases. This means that whenever system output is perceived as more diverse or of higher quality, the system is perceived as more useful. In addition, the racial identity of the participant of the participant seems to directly play a negative role in perceived utility of the system, and age directly influences the quality perceived. The SEM model also shows a positive relationship between perceived diversity and perceived quality, suggesting that a higher perceived diversity in the output influenced utility both directly and mediated through quality.

To investigate the influence of the mitigation conditions on perceived utility, a look can be taken at the mean factor scores, calculated using final the results of the CFA, for the conditions for the perceived system aspects in Figure6. What can be seen is that perceived utility is highest for the equal mitigation condition compared to the other conditions. This can be explained in most part by the higher perceived diversity for this condition, and due to the larger effect size of diversity compared to quality as visible in the SEM model 5. For perceived quality, it can be seen how the baseline condition has the highest score, which, despite the low perceived diversity of this condition, causes the perceived utility

of this condition to be not as low as the inverse condition, which has both a comparably low perceived diversity and quality. The effect of the high perceived quality, however, only in part mediates the stronger negative effect of low perceived diversity.

Fig. 6. Mean calculated factor scores for each mitigation condition, per perceived system aspect (top), and further explored per bias condition (bottom). Error bars indicate one Standard Error.

The differences in perceived diversity between conditions depends on the underlying biases in the mitigated output. The SEM model shows how both a male and a female bias have a negative effect on perceived diversity compared to the "both" gender bias condition. The biases can also interact with the intervention conditions. For the baseline condition, it can be seen in the SEM model how perceived diversity is only negatively affected by the baseline condition compared to the equal mitigation condition when the baseline images display male gender bias. However, for the inverse mitigation condition, a negative effect on the perceived diversity compared to the equal mitigation condition can be seen regardless of the gender bias underlying the mitigated images. Furthermore, regardless of intervention, the racial bias displayed in the mitigated images also influenced perceived diversity.

Figure 6 shows the mean factor scores for the interactions between the intervention and gender bias conditions. Compared to the both gender bias condition, the male and female conditions have lower average perceived diversity scores. In addition, the negative interaction effect between the male gender bias condition and the baseline intervention condition can be recognised in the low perceived diversity factor score.

The differences in perceived quality between conditions also partially depend on the underlying biases in the mitigated output. The SEM model shows how the female bias condition has a positive effect on perceived quality compared to the "both" gender bias condition. For the baseline mitigation condition, a positive effect on the perceived quality compared to the equal mitigation condition can be seen regardless of the gender bias underlying the mitigated images. However, for the inverse mitigation condition, the positive effect on perceived quality is completely moderated by the female bias condition, resulting on a negative effect on perceived quality compared to the equal mitigation condition. Furthermore, like for perceived diversity, the POC bias condition was found to positively influence perceived quality, compared to the white bias condition. In addition, the age of the participant was found to positively influence perceived quality.

Figure 6 also shows the mean factor scores for perceived quality and the interactions between the intervention and gender bias conditions. Compared to the both gender bias condition, the male and female conditions have similar scores. However, the negative interaction effect between the female gender bias condition and the inverse intervention condition can be seen in the relatively low perceived quality score. The general positive influence of the female gender bias condition, is especially evident in the graph for the baseline condition.

4.2 Response Model including surprise

Although perceived surprise was excluded as a factor, one of the questions loading on perceived surprise was found to be able to individually contribute to the SEM model. The question “The output had images I would not have thought to consider” was added to the model as an indicator variable, as it had a relatively high R^2 compared to the other questions (0.25, see Table 3) and low cross-loadings with the other perceived aspects. After an iterative process of carefully removing insignificant effects as described before, the resulting model is shown in Figure 7 with a good fit ($\chi^2(140) = 192.124$, p =0.002, CFI = 0.994, TLI =0.998, RMSEA = 0.029, 95% CI : [0.018, 0.039]). Although the variable does not by itself represent the perceived surprise factor, it will be considered as a representation of part of this factor.

The thick arrows in the SEM model indicate new significant effects and insignificant effects that were previously significant in the SEM model of Figure 5. What can be seen, is that like diversity and quality, the question has a positive effect on utility. In addition, perceived diversity is found to positively influence the variable relating to surprise, meaning that outputs that are perceived as more diverse will possibly be perceived as more surprising, or contain images that the user would not have thought to consider. The variable has, however, a negative effect on the perceived quality. This means that the perceived quality of the output is positively influenced if they are in line with the user’s expectations.

Again the differences in perceived the responses to the questions might depend on the underlying biases in the mitigated output. A noteworthy difference between the models of Figures 5 and 7 is that the effect of the female bias condition on perceived quality is no longer considered significant. In the original SEM, the female bias condition improved quality, which in turn improved utility. However, in the adjusted SEM, this effect has shifted to female bias negatively influencing the indicator variable for perceived surprise, which in turn decreases utility, but improves quality, which has a slight positive effect on utility. Furthermore, compared to the equal mitigation both gender condition, the baseline and inverse conditions do not have a direct influence on the indicator variable that can indirectly improve utility. However, in the female gender bias condition, there is an interaction effect with the inverse mitigation condition, resulting in an increase of the item for surprise. Lastly, the influence of the ethnic identity of the participant now also negatively affects the indicator variable, whereas the direct negative effect on perceived utility is smaller.

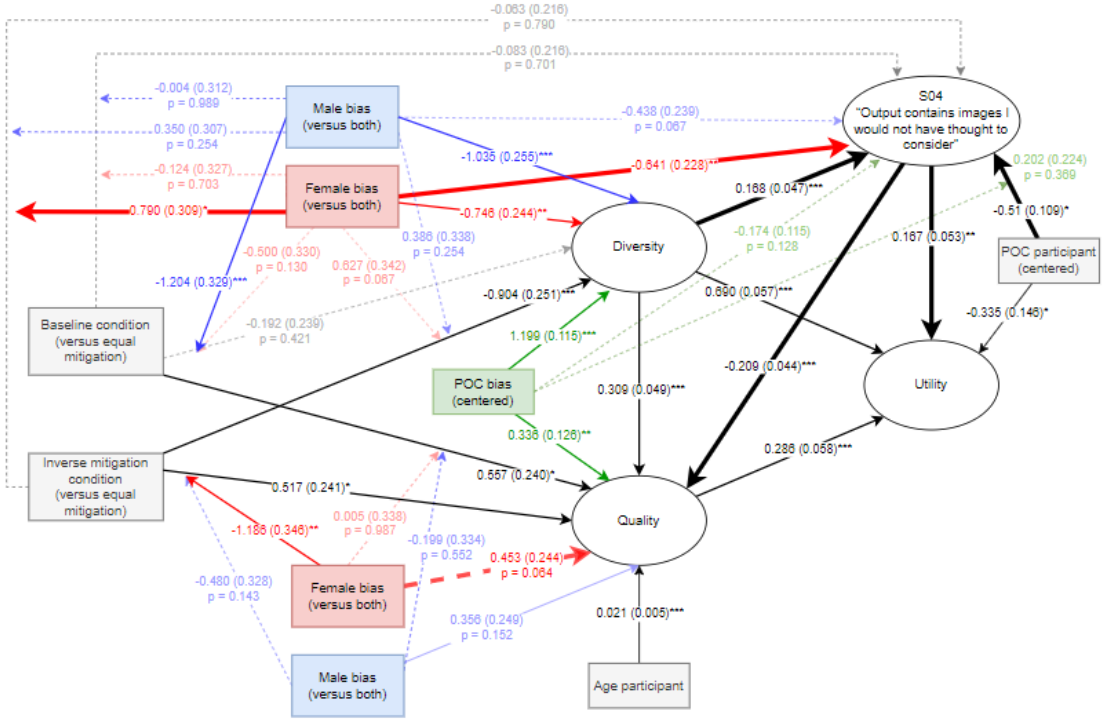

Fig. 7. SEM model including one of the questions previously loading on surprise. It shows the relations between the conditions and perceived system aspects. Thick arrows have been used to indicate new significant effects and insignificant effects that were previously significant in the SEM model of Figure 5. Arrows have their coefficients with standard error between brackets and p-values. If no p value is indicated, *** = p<0.001, ** = p<0.01, *= p<0.05. Dotted lines indicate relations with p>0.05.

5 DISCUSSION

This study sought to investigate how the implementation of a social bias mitigation strategy impacts the perceived effectiveness of an image generation system. Working upon, and going beyond, conventional evaluation methodologies auditing the outputs of a text-to-image model, the User-Centric Evaluation Framework [53] was used to do a novel user study to evaluate the perceptions of the system. Simulating a real-life professional scenario in the advertising industry, the influence of a promising mitigation technique on the utility of the tool in a creative ideation task was investigated.

5.1 RQ1: How do user's perceptions of diversity, quality, and surprise influence the perceived utility of the T2I system?

In line with research on effective ideation[54], perceived utility of the system was found to increase with perceived quality and diversity. In addition, perceived diversity was found to positively influence perceived quality. These effects underscore the importance of increasing perceived diversity of T2I system outputs. It can improve perceived utility of the system for an ideation task with diversity-related goals, as well as the quality of the output that was experienced. However, it also illustrates that the perceived quality of a T2I the system output might not be evaluated best by focusing on single output images in isolation. Such an evaluation cannot take into account the effect perceived diversity might

have in such systems. Because of the inherent generative variability[5] of such systems, users will always be exposed to a variety of images when interacting with the system, and thus a level of perceived diversity that can influence perceived quality and needs to be accounted for in evaluation, specifically in the context of social biases.

5.2 RQ2: How does a social bias mitigation strategy influence the perceived utility of the T2I system?

Overall, the effects of the equal mitigation method on perceived utility was found to be positive compared to the baseline and inverse mitigation condition. This effect could be attributed to how the equal mitigation method increased perceived diversity. However, this effect depends on the gender bias underlying the images. Not surprisingly, the equal mitigation was not found to outperform on perceived diversity compared to the baseline condition when there was no underlying single gender bias. However, when mitigating output with a female gender bias, perceived diversity did not increase compared to the unmitigated condition. A simple explanation for this might be that the mitigation was not sufficiently effective in mitigating the bias in these images and improving diversity. In line with this, Friedrich et al. [41] also find variance in the mitigation results, finding a higher variance for effectiveness of mitigating female gender-bias. That increasing perceived social diversity is a complex task, is also evident in the direct negative effects of underlying female and male gender biases on perceived diversity. Images with a female or male gender stereotype are not perceived equally diverse compared to when there's no underlying single gender bias, even when mitigated. This suggests that there are gender presentation differences of the underlying gender biases [7], that make these images perceived less diverse, even when mitigated and controlled for multiple factors such as the racial bias and the user's own gender identity.

The perceived quality of the system output also depends on the underlying gender biases in the images. When there is no strong female or male gender bias underlying the images, the direct effect of equal mitigation on perceived quality is negative, compared to the baseline and inverse mitigation conditions. This could suggest a negative effect of the steering that the mitigation does, which contradicts the findings by Brack et al. [52] of increased quality after semantic guidance. However, this negative effect might only appear when steering images in the direction of the opposite gender [44]. However, if steering into the opposite gender direction decreases perceived quality, this would mean that both in the female and male bias conditions there would be an expected moderated negative influence of the inverse mitigation on perceived quality. Only a when the underlying bias is female, a significant negative interaction effect is visible. Such an asymmetric effect, of underlying female bias and not male bias, was also seen in the direct effects of this bias on perceived quality, where underlying female bias positively influences perceived quality.

An adjusted SEM model including an indicator variable relating to surprise gave some further insights on the dynamics between perceived system aspects [54]. Perceived diversity positively influences the indicator, that in turn stimulates perceived utility and negatively influences quality. The findings suggests that although the surprising results might come at the cost of the perceived quality, they still aid the perceived usefulness in an ideation task. This aligns with one of Weisz et al.'s [5] design principles for generative AI is to design for imperfection. Whereas the output might be imperfect in that it is too novel or not what the user expected, this might still lead to a useful synergy. By introducing the indicator variable, the direct effect of underlying female bias on perceived quality also became insignificant. However, in turn, an underlying female bias was found to negatively influence the perceived surprise, which positively influenced the quality perceived. This suggests that system output with an underlying female bias will be perceived of higher quality compared to images without an underlying single gender bias, not because they are of higher quality in itself, but because they are in line with someone's expectations. This seems to align with the findings of Flory et al. [56], who found that search results were perceived more relevant in stereotypical female category, when

the female bias was expressed. As such, a displayed bias confirming that of the user might not aid in ideation, but be perceived of higher quality. Lastly, the power of interaction effect of female bias with the inverse mitigation on perceived quality was also smaller, while an additional interaction effect of female bias with the inverse mitigation going through surprise was found. This suggests that the inverse mitigation resulted in an output the user would not have had considered, positively influencing the perceived utility of the system.

5.3 RQ3: How does the social bias context of the unmitigated images, as well as the personal context of the user, influence the perceived utility of the T2I system?

As evident from the previous paragraphs, the effect of the baseline and mitigation conditions on perceived utility could not be addressed without addressing the gender biases underlying the images. Other contextual factors were also found to influence perceived utility, such as the racial bias in the underlying images. Besides perceived diversity, underlying POC racial bias in the images was also found to directly positively influence perceived quality, compared to the white racial bias setting. Again, this might be due to complex presentation differences of the underlying racial biases [7, 39]. Because the semantic guidance method [52] only edits the concept that is steered towards, no interaction effects were found between the mitigation conditions and the underlying bias. However, brief qualitative inspection of the effects of mitigation in the "Cook" images, as can be seen in B, shows editing gender biases can accidentally result in the editing of the racial appearance too.

In contrast to the finding of Metaxa et al. [34] that the user's gender identity matters in perception of occupational biases, such interactions were not found. However, whether the participant identified their ethnicity as white or not was found to directly influence perceived utility. This indicates that, even in the context of mitigating gender bias, controlling for racial biases in the outputs, and focusing on enhancing perceived gender diversity, someone's ethnicity, rather than gender, is a crucial factor in influencing how they perceive the overall utility of these measures. Part of this effect could be explained through the adapted SEM model, by illustrating how someone's ethnic background influences what the participant would have considered or not. The system output was less likely to contain surprising images to people not identifying as white.

Lastly, older participants also perceived the quality as higher compared to younger ones, however, this could not be explained through a lower perceived level of diversity, surprise, or through a positive effect directly on utility. Perhaps, older people are less critical on the quality of images generated. This could come because of a lower affinity with such systems and their capabilities . Although care was taken to provide an example in the scenario description (see Fig 3), users more familiar with image generation technology might be aware of its far reaching capabilities, and might not be impressed by the quality of images produced by an older foundation model like SD v1-5.

5.4 Limitations

Because this was a first attempt of applying a User-Centric Framework [53] to the evaluation of T2I systems and their social bias mitigation, the study is not without its limitations. This study sought to measure the influence of different system aspects on perceived utility in an ideation task, including the effect of novelty through the surprise factor. Due to poor validity the surprise factor was later dropped, however the evaluation of its influence helped demonstrate the extent to which the question captured the impact of user's expectations on the perceived quality and utility of the model. When analysing the surprise factor within the model as an indicator variable, the power of these findings is limited. However, in the context of the utility for overcoming social biases, the other questions expected to load on surprise likely captured elements of surprise that might not relate to it. Similarly, two other survey items were

also subsequently dropped, the items on perceived quality and utility. Firstly the item on perceived quality although often related to quality or accuracy of T2I systems, it might capture a different dimension of quality, alignment, rather than fidelity [64]. Lastly, a question expected to load on perceived utility, expectantly because in contrast to the other questions it did not directly relate to the task.

The experimental setup used has its strengths and weaknesses for evaluating mitigated image generation systems. The mixed design allowed for direct comparison within participants, as well as capturing nuanced differences between participants, by using images with the same seed. Because images only differed in mitigation condition, the influence of the mitigation was directly comparable. However, the use of pre-generated images and system screenshots to mimic a real local system, is unable to provide users with control to setup the system as desired, and might not capture the influence of the generative variability that T2I systems offer [5], nor the dynamics that the Fair Diffusion mitigation [41] offers by allowing for control during deployment. This would empower the user to design their own fair instructions, fully adapted to their context.

Striving to maintain the experiment similar to the experience of simulating a real T2I system, the selection of occupations was solely done based on labour statistics and previous findings of auditing studies. In addition, per occupation, only three batches of 9 baseline images were generated out of which one was selected based on portrayal of the expected underlying biases. The outputs after mitigation were only inspected for possible harmful content. However, beyond this limited qualitative evaluation, no automated metrics for diversity or quality have been used in the selection of the images to further verify alignment with previous research. While this approach ensures a degree of experimental authenticity within the constraints of control or variability in results, it also indicates that further investigation is required to understand the precise impact and to directly compare 'objective' metrics with metrics of perceived diversity and quality.

Further generalisation of the findings also needs to be verified. To begin with, all participants of the user experiments were based in the US, to ascertain a level of alignment between the gender and racial bias present in the model and the US labour statistics. This was to build upon existing research and avoid large differences in stereotypes that participants hold and the model displays. However, this research has demonstrated that one's ethnic identity and expectations can influence the perception of the system, meaning that it should be explored in a more global setting. Future research should also go beyond occupation biases, beyond gender and race, and have a closer look at the intersectionality. The mitigation method also allows findings to be verified in any framework employing classifier free guidance. It was shown how its effect extend to various models [52], but because biases differ between models comparing this might still be an interesting case.

Lastly, this study acknowledges the limitations inherent in using binary categories for gender (male and female) and race and ethnicity (POC vs. white) when analysing social biases in text-to-image models. Although such an approach is common in related research, these categories tend to oversimplify complex identities. Both gender and race encompass a spectrum far broader than binary options. By focusing solely on these binary divisions, the methodology of this research risks perpetuating stereotypes and failing to represent non-binary and racially diverse individuals. Moreover, it overlooks the intricate interplay of various types of biases and the fact that many attributes cannot be visually deduced from generated images. This limitation underscores the necessity for more inclusive and nuanced methods in future research. While this work marks a step towards understanding social bias perception in image generation, it also emphasises the ongoing need for more comprehensive and ethical approaches to tackle these complex, multifaceted social constructs.

6 CONCLUSION

This study has set a first step in the direction of T2I system evaluation of social bias mitigation, that goes beyond classifying outputs. It has highlighted the need for similar approaches, that take into account the complex dynamics behind the perceived usefulness of a T2I system for a real-life task. This is a novel approach to the measurement of the qualities of a T2I system in real-life applications, inspired by Recommender Systems research [53]. By performing a user-experiment allowing for direct comparisons between images and two different gender mitigation conditions, controlling for a possible effect of racial bias in the images, and taking into account user's gender and racial identity, the findings of the evaluation of the mitigation method proposed by Friedrich et al. [41] were extended.

Increased perceived diversity of the system was found to relate to a higher perceived utility, as well as perceived quality. However the application of these mitigation strategies also has effects on the perceived quality. The effectiveness of the mitigation strategies to raise perceived diversity, while maintaining perceived quality, varies strongly through the underlying biases associated to the images. Within the context of ideation and the capacity for this tool to provide a creative output for real-life application, the effect of perceived surprise was explored. Although not applicable to a general model, through single item as an indicator value it allowed to explore the connection between surprise and perceived utility and quality. Suggesting that if output is in line with user's considerations, this is beneficial for the quality perceived, but not for utility in the ideation task. Furthermore, the identity of the user was found to directly influence perceptions. This should be addressed through further research that examines different user identities and context and its effect in perceptions through varying underlying biases.

The use of pre-seeded images and system screenshots to create the user-experiment, takes away the ability to control from the users within the T2I tools and mitigation methods. In a real-life setting, users can setup their environment and tools according to their requirements, including personal and professional underlying biases and a wide range of diversity objectives. Further research that explores the use of tools in a free environment with fine-tuning capabilities is warranted to further develop the findings made in this study.

This study notes that harmful racial and gender biases can have a detrimental impact on personal and professional contexts, and that biases in T2I might persist due to the inherently biased nature of large internet-crawled datasets, as well as varying perceptions of fairness throughout communities. Therefore research must conduct findings that allow for new systems to be applied in any societal context while keeping the users from being unaware of the promotion of underlying racial and gender biases.

REFERENCES

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. arxiv. *arXiv preprint arXiv:2112.10752*, 2021.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2023.

[3] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794, 2021.

[5] Justin D Weisz, Michael Muller, Jessica He, and Stephanie Houde. Toward general design principles for generative ai applications. *arXiv preprint arXiv:2301.05578*, 2023.

[6] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759*, 2022.

[7] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. *arXiv preprint arXiv:2302.03675*, 2023.

[8] Fabrizio Santoniccolo, Tommaso Trombetta, Maria Noemi Paradiso, and Luca Rollè. Gender and media representations: A review of the literature on gender stereotypes, objectification and sexualization. *International Journal of Environmental Research and Public Health*, 20(10):5770, 2023.

[9] Erica Scharrer, Srividya Ramasubramanian, and Omotayo Banjo. Media, diversity, and representation in the us: A review of the quantitative research literature on media content and effects. *Journal of Broadcasting & Electronic Media*, 66(4):723–749, 2022.

[10] Robert E McDonald, Debra A Laverie, and Kerry T Manis. The interplay between advertising and society: an historical analysis. *Journal of Macromarketing*, 41(4):585–609, 2021.

[11] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[12] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[13] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[14] Judith Lorber, Susan A Farrell, et al. *The social construction of gender*. Sage Newbury Park, CA, 1991.

[15] Audrey Smedley and Brian D Smedley. Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. *American psychologist*, 60(1):16, 2005.

[16] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023.

[17] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[18] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[19] Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. Reducing the carbon impact of generative ai inference (today and in 2035). In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, pages 1–7, 2023.

[20] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[22] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[23] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[24] Zhangyin Feng, Runyi Hu, Liangxin Liu, Fan Zhang, Duyu Tang, Yong Dai, Xiaocheng Feng, Jiwei Li, Bing Qin, and Shuming Shi. Emage: Non-autoregressive text-to-image generation. *arXiv preprint arXiv:2312.14988*, 2023.

[25] Cheng Li, Yali Qi, Qingtao Zeng, and Likun Lu. Comparison of image generation methods based on diffusion models. In *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, pages 1–4. IEEE, 2023.

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[29] Stable diffusion v1-5, 2022. URL https://huggingface.co/runwayml/stable-diffusion-v1-5.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] J Alammar. The illustrated stable diffusion, 2022. URL https://jalammar.github.io/illustrated-stable-diffusion/.

[32] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

[33] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[34] Danaë Metaxa, Michelle A Gan, Su Goh, Jeff Hancock, and James A Landay. An image of society: Gender and racial representation and impact in image search results for occupations. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

[35] Swagatika Dash and Yunhe Feng. Fairness in image search: A study of occupational stereotyping in image retrieval and its debiasing. *arXiv preprint arXiv:2305.03881*, 2023.

[36] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*, 2022.

[37] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022.

[38] Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. Is the elephant flying? resolving ambiguities in text-to-image generative models. *arXiv preprint arXiv:2211.12503*, 2022.

[39] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

[40] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

[41] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

[42] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. *arXiv preprint arXiv:2303.08084*, 2023.

[43] Marc Cheong, Ehsan Abedin, Marinus Ferreira, Ritsaart Willem Reimann, Shalom Chalson, Pamela Robinson, Joanne Byrne, Leah Ruppanner, Mark Alfano, and Colin Klein. Investigating gender and racial biases in dall-e mini images.

[44] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*, 2023.

[45] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

[46] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[47] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 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*, pages 3819–3828, 2015.

[48] Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*, 2019.

[49] Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. A friendly face: Do text-to-image systems rely on stereotypes when the input is under-specified? *arXiv preprint arXiv:2302.07159*, 2023.

[50] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*, 2022.

[51] Zhixuan Liu, Youeun Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark (ccub) dataset. *arXiv preprint arXiv:2301.12073*, 2023.

[52] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[53] Bart P Knijnenburg and Martijn C Willemsen. Evaluating recommender systems with user experiments. In *Recommender systems handbook*, pages 309–352. Springer, 2015.

[54] Jami J Shah, Steve M Smith, and Noe Vargas-Hernandez. Metrics for measuring ideation effectiveness. *Design studies*, 24(2):111–134, 2003.

[55] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. *Recommender Systems Handbook*, pages 1–35, 2021.

[56] Jeffrey Flory, Andreas Leibbrandt, Olga Shurchkov, Olga Stoddard, and Alva Taylor. Consumer preferences for diversity: A field experiment in product design. *Available at SSRN*, 2022.

[57] Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. Do perceived gender biases in retrieval results affect relevance judgements? In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 104–116. Springer, 2022.

[58] Elaine Swan. Commodity diversity: Smiling faces as a strategy of containment. *Organization*, 17(1):77–100, 2010.

[59] Michael D Ekstrand, F Maxwell Harper, Martijn C Willemsen, and Joseph A Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 161–168, 2014.

[60] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems–a survey. *Knowledge-based systems*, 123:154–162, 2017.

[61] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.

[62] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: A survey. *arXiv preprint arXiv:2307.04644*, 2023.

[63] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User modeling and user-adapted interaction*, 22:441–504, 2012.

[64] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023.

[65] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 157–164, 2011.

[66] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. How serendipity improves user satisfaction with recommendations? a large-scale user evaluation. In *The world wide web conference*, pages 240–250, 2019.

[67] davisking. Github - davisking/dlib: A toolkit for making real world machine learning and data analysis applications in c++, May 2023. URL https://github.com/davisking/dlib?tab=readme-ov-file.

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[69] bls. Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity. *Labor Force Statistics from the Current Population Survey*, 11, 2022.

[70] Abhishek Mandal, Susan Leavy, and Suzanne Little. Dataset diversity: measuring and mitigating geographical bias in image search and retrieval. In *Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing*, pages 19–25, 2021.

[71] 2023. URL https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F.

[72] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[73] Yves Rosseel. lavaan: An r package for structural equation modeling. *Journal of statistical software*, 48:1–36, 2012.

[74] R Developement Core Team. R: A language and environment for statistical computing. *(No Title)*, 2010.

[75] Barbara M Byrne. *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. routledge, 2013.

Appendices

A USER EXPERIMENT SCENARIO AND TASK DESCRIPTION

Scenario description

Imagine you have a creative role at a marketing agency. You are asked to **help create an image for a campaign for a job agency.** With the image you will design the job agency hopes to attract people looking for a job to their website.

One of the brand values of the job agency is **diversity, equity, and inclusion (DE&I)**. These principles are aimed at welcoming and supporting individuals from diverse backgrounds and life experiences. Diversity ensures representation of different backgrounds and perspectives, Equity aims for fairness and equal opportunities, and Inclusion focuses on making everyone feel valued and heard. Because of these values, the agency hopes to **reach a diverse set of people** with their job advertising. Currently, the agency is mostly interested in **avoiding gender bias** in job advertising.

To create the image for the advertisement, you will use a **text-to-image generation system**. The system uses **artificial intelligence** to generate images based on a text prompt you give to it. An image of such a system can be seen in Figure 1 below. Here as an example the prompt "A photograph of Obama eating an apple" is used to output various pictures.

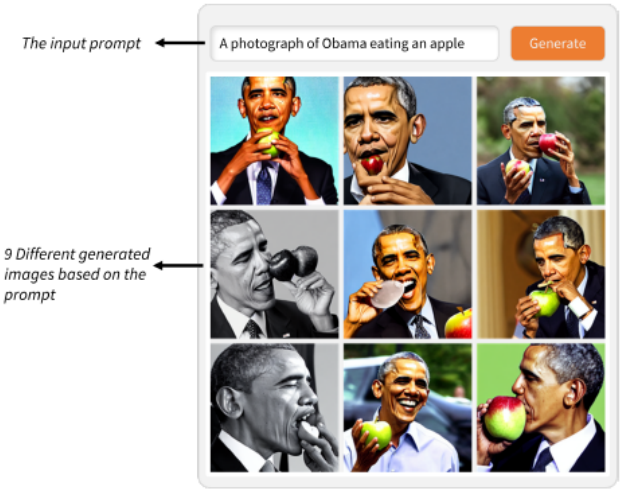

Figure 1. An example screenshot of the system

Figure 1 shows how the system generated **9 random images for the same prompt**: "'A photograph of Obama eating an apple". You can also see how the system can also give odd outputs, such as deformations of hands and misplaced objects. Because of this, rather than generating a picture that is perfect for the campaign, you decide to use the system for **brainstorming**, to come up with ideas for the **main person featuring in the campaign.** You begin by generating the image of a face of a person with a certain profession.

For now, **your focus is on the people in the generated images.** Details such as adding text to the image or resizing it are not of importance yet.

Task description

Next, you will see a screenshot similar to Figure 1 of the system you are using. It will show the prompt you have used as input, as well as the outputs the system has generated as a result. You will be asked to **evaluate the system and its outputs based on this screenshot.** In total, you will be asked to review 3 system screenshots.

Before evaluating these screenshots, there will first be a short comprehension check to see if you have carefully read and understood the assignment.

Fig. 8. User experiment scenario and task description

B IMAGES USER EXPERIMENT

The images below were displayed in the user experiment as part of the system screenshot as displayed in A. The prompt: "A photo of the face of a [occupation]" was used to generate the images under the three mitigation conditions.

Fig. 9. Nursing Assistant - Baseline condition

Fig. 10. Nursing Assistant - Equal mitigation condition

Fig. 11. Nursing Assistant - Inverse mitigation condition

Fig. 12. Dental hygienist - Baseline condition

Fig. 13. Dental hygienist - Equal mitigation condition

Fig. 14. Dental hygienist - Inverse mitigation condition

Fig. 15. Security Guard - Baseline condition

Fig. 16. Security Guard - Equal mitigation condition

Fig. 17. Security Guard - Inverse mitigation condition

Fig. 18. Electrician - Baseline condition

Fig. 19. Electrician - Equal mitigation condition

Fig. 20. Electrician - Inverse mitigation condition

Fig. 21. Cook - Baseline

Fig. 22. Cook - Equal mitigation

Fig. 23. Cook - Inverse mitigation

Fig. 24. Producer - Baseline

Fig. 25. Producer - Equal mitigation

Fig. 26. Producer - Inverse mitigation