Investigating gender and racial biases in DALL-E Mini Images

Marc Cheong[a,d,1], **Ehsan Abedin**[a], **Marinus Ferreira**[b], **Ritsaart Reimann**[b], **Shalom Chalson**[c], **Pamela Robinson**[c], **Joanne Byrne**[d], **Leah Ruppanner**[e], **Mark Alfano**[b], **and Colin Klein**[c]

[a]School of Computing and Information Systems, University of Melbourne, Parkville VIC 3010, Australia; [b]Philosophy, Macquarie University, Macquarie Park, NSW 2109, Australia; [c]Philosophy, Australian National University, Canberra ACT 2601, Australia; [d]Centre for AI and Digital Ethics, University of Melbourne, Parkville VIC 3010, Australia; [e]School of Arts, University of Melbourne, Melbourne, Parkville VIC 3010, Australia

This manuscript was compiled on April 5, 2023

Generative artificial intelligence systems based on transformers, including both text-generators like GPT-3 and image generators like DALL-E 2, have recently entered the popular consciousness. These tools, while impressive, are liable to reproduce, exacerbate, and reinforce extant human social biases, such as gender and racial biases. In this paper, we systematically review the extent to which DALL-E Mini suffers from this problem. In line with the Model Card published alongside DALL-E Mini by its creators, we find that the images it produces tend to represent dozens of different occupations as populated either solely by men (e.g., pilot, builder, plumber) or solely by women (e.g., hairdresser, receptionist, dietitian). In addition, the images DALL-E Mini produces tend to represent most occupations as populated primarily or solely by White people (e.g., farmer, painter, prison officer, software engineer) and very few by non-White people (e.g., pastor, rapper). These findings suggest that exciting new AI technologies should be critically scrutinized and perhaps regulated before they are unleashed on society.

gender bias | racial bias | algorithmic bias | generative AI | DALL-E Mini

Generative artificial intelligence systems based on transformers have recently entered the popular consciousness. The more popular ones include GPT-3 (Generative Pre-trained Transformer 3) and more recently ChatGPT which are able to generate textual content based on an input prompt; and DALL-E[*] 2 which is similarly able to generate images with a similar prompt.

These generative systems are based on transformer systems: complex neural network algorithms which, in a nutshell, "learns context and thus meaning by tracking relationships in sequential data" (1). These systems depend on large data models—akin to their "vocabulary"—which have been trained on a large collection of images, text, and the relations between them, over many iterations. The end-user merely has to provide a prompt as input to the system, which then uses its model to generate candidate outputs that closely match the prompt.

For example, given the prompt, "In a paragraph, what are the benefits of AI image generation systems?", ChatGPT's response is rather convincing: "These systems can create realistic images of objects, scenes, and people that do not exist in reality, which can be used for creative applications such as generating new designs for clothes, furniture, and other products. They can also reduce the cost and time of producing new images, create personalized images for individuals, aid in medical imaging, and create realistic images and animations for video games and the film industry."[†].

In the domain of image generation, one of the current state-of-the-art technologies, as of time of writing, is DALL-E 2, owned and operated by the OpenAI consortium. Its open-source derivative, DALL-E Mini (2) is widely available (via its Craiyon.ai web app), is easy-to-implement (with sample programming code provided freely for reuse), and is able to generate images with virtually no cost or barrier to entry. Its image generation capabilities are not as extensive as DALL-E, but the entire model has the advantage of being readily deployed on any modern computer or cloud-based programming environment (such as Google Colab) in a matter of minutes. To better understand how generative AIs—specifically DALL-E Mini—work, we offer a birds-eye-view of the technology here.

A Primer on Generative Technologies. As DALL-E mini shares characteristics with systems including DALL-E (OpenAI) and GPT-3 (Brown et al., 2020), which DALL-E is based upon, it will suffice to give a general overview of the technology.

First, an image model is trained on a large collection of images with associated captions. For DALL-E Mini, a dataset of over ~15M images used in machine learning research (3, 4) is passed through an encoder called VQGAN (5). These datasets are *de rigeur* in the machine learning community as they allow

[*] Stylised DALL·E; it is based on GPT-3 but produces images instead of text as outputs.

[†] Edited from prose generated with ChatGPT Feb 13 Version. Free Research Preview.

Significance Statement

DALL-E Mini, an example of a Generative AI system, is able to produce images and artwork based on prompts given by the user. However, as with many AI systems, it has to learn about art from somewhere—and that 'somewhere' is a large collection of images, text, and the relations between them created by humans. Thus, it encapsulates human and societal biases, including gender bias and racial bias. This work aims to measure the degree of gender and racial bias that DALL-E Mini may be vulnerable to, by comparing how it 'perceives' certain occupations with the reality of who is working in those occupations.

MC is involved in programming and drafting the manuscript. MC, EA, MF, MA, RR, and CK contributed to the design of the study. EA and MF wrote the codebook. MC, RR, EA, JB, and LR contributed to the literature review. MC, EA, MF, MA, RR, CK, SC, and PR contributed to the coding of the images in the dataset. EA did the majority of statistical analysis. All authors contributed to editing the manuscript and analysis of results.

No competing interests to declare.

[1] E-mail: marc.cheongunimelb.edu.au

for standardised experimentation; images within are taken from sources such as Flickr.

This in effect "turns images into a sequence of tokens" where the images' caption/description text are "encoded through a BERT encoder" (4). Both sets of encoded features (tokens) are processed by the "BERT decoder, which is an auto-regressive model whose goal is to predict the next token" (4). In short, this final step is used to associate the features (tokens) of each image with the features of each description based on their statistical likelihood.

When the user presents DALL-E mini with a prompt, the BERT encoder works on the text as before. Mirroring the training step, the text features (tokens) are used to predict what image features are likely to be associated with them. VQGAN is then used, albeit in a mirrored fashion, to decode these image features into actual graphical representations (3, 4).

The models based on the aforementioned technologies are constructed based on a large assemblage of human input: for example, an image generation system would learn from a large collection of input images to infer graphical properties related to certain concepts: e.g., what makes the image of a doctor (scrubs, stethoscope) different from a chef (cooking apron, kitchen equipment). These concepts are operationalized as a vast series of correlations: for each token, it is encoded as a list of which each entry measures the extent to which it is likely to co-occur with each other token, taking into consideration its associated linguistic contexts and distribution within a unit of text (6). So, when the system sees 'doctor', it makes 'syringe' much more likely to appear than 'spatula'. We are not the first to point out that the data used to train such systems are not free from—and indeed, are essentially dependent on—human bias. Furthering the example, if the majority of the images we use to train an image generation system are of white men in the medical profession, these systems will unavoidably pick up a correlation between these features and being linked to the token 'doctor', since that is simply how these tokens function within the system.

The Problem with Generative AI. As can be seen in recent literature in the field of AI ethics and the impact of technology on society (7, 8), such systems are rife with systemic flaws that have origins in the data used to train and build them, and are manifest as emergent behavior. Of particular concern is the issue of *bias*[‡] (9–11)—the propensity of such systems to reflect, entrench, and reinforce harmful stereotypes and prejudice that exist in society writ large (12).

Despite initial public perception that AIs are unbiased (Bryson, as cited in (13)), far from realizing the espoused ideal of impartiality, AI bias is both pervasive and pernicious: implicating everything from unequal access to health care (14, 15) and education (16, 17); to reduced employment prospects (18, 19) and racially skewed rates of (re)incarceration (20, 21). Add to this list increased risk of medical misdiagnosis (22, 23), unequal financial opportunity (24), and greater vulnerability to self-driving cars (25), and we begin to get a sense of just how far reaching the effects of AI bias are.

[‡] It is worth noting that the term 'bias' has different connotations in computing/mathematics; we qualify our current use of 'bias' as "prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair" (per the New Oxford American Dictionary).

Landmark cases on racial and gender bias in extant AI systems include the following: Amazon's hiring AI, which 'reads' CVs to determine an ideal candidate, was found to be gender-biased (26, 27); Google algorithms for search engines, photo tagging and ad placement were found to be racially biased (28–31); and systems that purportedly determine criminal risk of recidivism and crime patterns arguably reproduce racist biases (32–34).

Note that many of these *black-boxed* systems are inherently technologically complex, and therefore, these behaviors cannot merely be "switched off" at the touch of a button. To ameliorate the harms caused, an entire system may need to be decommissioned (in the case of Amazon's hiring system), or a stop-gap fix patched (in the case of Google's racist photo search algorithm).

AI bias also manifests more subtly in seemingly benign generative systems such as DALL-E Mini. The authors of DALL-E Mini, based on their ongoing evaluation (4, 35) acknowledge the inherent limitation of the technology: "Occupations demonstrating higher levels of education ... or high physical labor... are mostly represented by white men. In contrast, nurses, secretaries or assistants are typically women, often white as well."

They further highlight in their *Model Card* (36)—a report on the limitations and dangers of the models—that "initial testing demonstrates that they may generate images that contain negative stereotypes against minoritized groups." (37).

Potential implications of biases in visual representations of professional roles raise the possibility of an AI-mediated feedback loop (38): social biases embedded in generative models *encourage* biased decisions by human users, which in turn further *entrenches* those biases, both in the system and society at large. De-Arteaga et al.(39) also raise concerns about the *interdependency* of different models: what would happen if the results produced by one generative model become or influence the data used by another? Unsurprisingly, by investigating classifiers for occupational biographical profiles (bios), they find that subsequent generations of classifiers become progressively more gender-biased.

To further our inquiry, we turn to extant literature for analyses of racial and gender bias in similar generative systems. In their analysis of *minDALL-E* and *ruDALL-E-XL*, Cho et al. (40) find that when prompted with race- and gender-neutral terms, both algorithms return racialized and gendered output: typically coupling women and minority groups to menial work while reserving high status occupations for white men. In the same vein, Steed and Caliskan (41) found "racial, gender, and intersectional biases" in pre-trained image representation models.

These systems are reflecting back the statistically-dominant social group in each of these positions which undermines the nuance across occupations and deteriorates work being done to raise visibility of marginalized and minoritized groups within heavily-skewed industries. In essence, what DALL-E gains in speed and image generation efficiency, it loses in precision and nuance. And, for any group outside the socially dominant groups, this reinforces historical bias and marginalization. As such, gaining a clearer understanding of the extent to which multi-modal generative models are biased, what sorts of biases they perpetuate, and who suffers most at the hands of biased representation is of critical import (41, 42).

In this spirit, we seek to investigate the biases found in DALL-E Mini in a systematic fashion, when presented with prompts for a given occupation. The basic idea is this: if we were to ask DALL-E Mini to represent a doctor, we would expect the graphical representations of scrubs, a stethoscope, or the existence of a hospital, to be helpful discriminating characteristics (which will not be found in other careers such as chef or reporter). However, if the system thinks that 'doctor' indicates to the same or stronger degree with 'white man'— and if we are able to quantify how much this correlation differs from the *actual* labor demographics of the medical profession— we are then able to quantify a measure of bias in DALL-E Mini.

Results

We started with a dataset of DALL-E Mini created images (10 images $\times$ 105 occupations = 1,050 total), partitioned into five subsets, each of which were randomly assigned to a subset of the authors to code. Full details on the image generation process and technical parameters are in ***Materials and Methods***. A total of 6,900 coded data points were produced from this initial set of images.

The codebook consists of two independent dimensions: perceived gender of human figures in an image (man, woman, or indistinct) and perceived racial identity of the aforementioned figures (white or non-white). The proportions of gender and race for each career were determined by considering the consensus reached among the three coders. If at least two agreed on either gender or race, the image was assigned to that category. Otherwise (e.g., one said man, another said woman, and the third said indistinct), the record was excluded from the analysis.

The Fleiss multirater kappa (Table 1) results from the coding process varied depending on the dimension, but overall showed acceptable or high levels of reliability.

Table 1. Fleiss's multirater kappa for gender and race determination

Coded Subset	1	2	3	4	5
Gender: Man	0.86	0.87	0.83	0.96	0.88
Gender: woman	0.88	0.94	0.93	0.95	0.81
Gender: Indistinctive	0.56	0.64	0.66	0.64	0.58
Race: White	0.75	0.71	0.64	0.73	0.79
Race: Non-white	0.37	0.29	0.35	0.50	0.25
Overall	**0.73**	**0.76**	**0.73**	**0.79**	**0.74**

We then compare the proportion of per-occupation genders and races coded from our sample to the real-world distribution as found in the U.S. Bureau of Labor Statistics (43).

As part of this comparison, we removed occupations that were categorized as indistinct (from our coding), occupations from our dataset which form an archetype or superset of several occupations (such as "civil servant" or "business person"), and occupations that could not be located in the labor statistics (such as "lexicographer").

The distributions of the final list of 67 occupations and their corresponding real-world labor statistics are illustrated in Figures 1 and 2. (The complete data tables, as well as the distribution of genders in the labor force per occupation, are provided in the ***SI Appendix***).

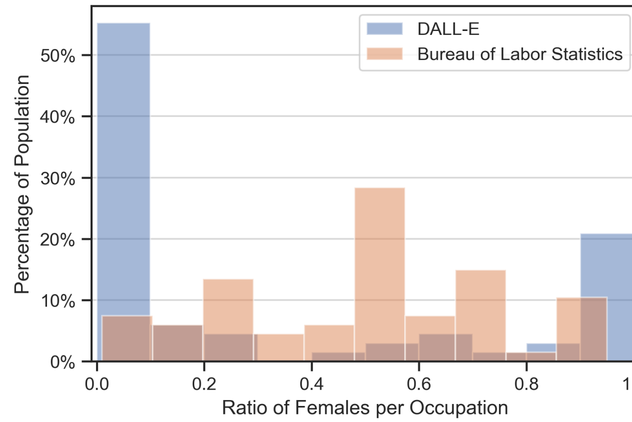

Fig. 1. Distribution of coded genders in our DALL-E dataset (in blue) versus actual baseline distributions per the Bureau of Labor Statistics (in orange). The vertical axis represents the percentage of the population within a group; while the horizontal axis indicates the ratio of women per occupation: 0.0 indicates that there are no women while 1.0 indicates that all of them are women.

As can be seen in Figure 1, the DALL-E Mini-generated images have a bimodal distribution - either completely men (left blue bar, i.e., proportion of women, at 0.00), or completely women (right blue bar). Compare this with the real-world distribution based on labor statistics (in orange). If DALL-E Mini were representative of the real-world gender distribution, the patterns we observe should be roughly the same, or, at the very least, symmetrical but non bimodal. To quantify the significance of the differences between DALL-E Mini's 'worldview' versus the real-world labor statistics, we conducted an independent samples t-test in IBM SPSS Statistics 28. To do so, we first made two grouping variables, Group 0 representing our coded DALL-E images, and Group 1 the official labor statistics. Regarding the gender difference between these two samples, our results show a statistically significant ($t = -2.88$, $p < 0.005$) difference between means: 0.318 for Group 0/DALL-E Mini and 0.489 for Group 1/labor stats.

Similarly, in Figure 2, the DALL-E Mini-generated images are overwhelmingly coded as containing White persons (right blue bar, i.e., proportion of White people at 1.00) around 85.07%. In contrast, the occupation with the *lowest* representation of images coded to be White (0.50) is (*rapper*). In other words, the DALL-E Mini images lack the nuanced distribution which is to be expected in real-world labor statistics, i.e., $\sim 55 - 96\%$ workers identified as White based on our occupational descriptors. Our findings from t-test, regarding the race difference between these two samples, again indicate a statistically significant difference ($t = -9.65$, $p < 0.005$) between the means of the two groups: 0.958 for Group 0/DALL-E Mini, 0.798 for Group 1/labor statistics.

Discussion

When we compare the occupations that DALL-E Mini represents as most gender-imbalanced, we find several stereotypes that are replicated – or entrenched – by this generative AI. This is most evident at the respective maxima. Thus, we analyze the list of occupations at each end of the bimodal distribution (i.e., either all men, or all women) in our DALL-E Mini dataset and compare them with actual labor statistics,

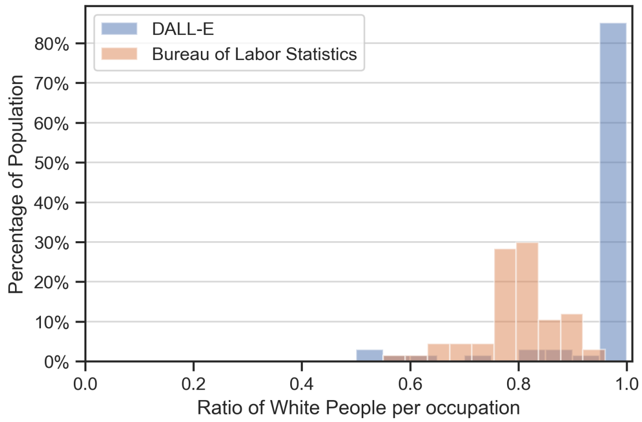

Fig. 2. Distribution of coded races in our DALL-E dataset (in blue) versus actual baseline distributions per the Bureau of Labor Statistics (in orange). The vertical axis represents the percentage of the population within a group; while the horizontal axis indicates the ratio of white people per occupation: 0.0 indicates that there are no white people while 1.0 indicates that all of them are white.

in Table 2.

Table 2. Highly-gendered occupational stereotypes in DALL-E Mini

DALL-E Mini versus Labor Statistics	Labor Statistics: high female representation	Labor Statistics: low female representation
DALL-E Mini high female representation	secretary, hairdresser, makeup-artist, receptionist, dietitian	salesperson, newscaster, newsreader, singer
DALL-E Mini low female representation	waiter, baker, accountant, biologist, poet, judge	pilot, builder, miner, electrician, plumber

When we consider the table's diagonal, we see the stereotypes of gendered work perpetuated. DALL-E Mini assumes that careers which are exclusively women include salesperson and singer, whereas the real-world statistics tell us otherwise: salespersons are fairly balanced ($\sim 49\%$ women), and singers have $\sim 26\%$ women. By contrast, roles such as biologist and judge are assumed by DALL-E mini to be predominantly men when in fact the actual statistics are $\sim 58\%$ and $\sim 56\%$ women, respectively. This is a reflection of occupational gender bias, a phenomenon documented in the sociological, psychological, and computing literature (44–48).

Similarly, DALL-E Mini is also likely to perpetuate racial bias in the images it generates. As mentioned in ***Results***, DALL-E Mini's 'worldview' is that almost all occupations are made up of White people. The exceptions are pastor, spokesperson, and rapper, where DALL-E Mini overestimated the racial balance of the workforce ($50\% \pm 10\%$, compared to the real-world average of $\sim 80\%$).

The findings above echo the DALL-E Mini Model Card (37) as discussed in the ***Introduction***. These results could be interpreted as the proverbial 'canary in the coalmine': alerting us to *downstream* consequences of social biases embedded in such generative AI systems (38, 42). As we have also observed, our results on race and gender bias in DALL-E Mini echo issues found in text-generation AIs and word embeddings (39, 41, 49–52).

Table 3. Racial occupational stereotypes in DALL-E Mini.

DALL-E Mini versus Labor Statistics	Labor Statistics: higher White representation	Labor Statistics: balanced representation
DALL-E Mini higher White representation	pilot, farmer, painter, electrician	doctor, physician, prison officer, chef, software engineer
DALL-E Mini balanced representation	pastor, spokesperson, rapper [Note 1]	N/A [Note 2]

Notes: [1] Although DALL-E Mini represents White and non-White groups fairly (both spokesperson and rapper at $\sim 50\%$, conversely, labor statistics indicate that the proportion of Whites are approximately $\sim 80\%$. [2] There is no occupation in our list that has balanced representations ($50\% \pm 10\%$) for *both* DALL-E Mini and real-world distributions.

DALL-E Mini may be capturing the racial and gender composition of the images on the Internet which do not replicate the statistical distribution within the labor market (43). Again, this points to the fact that these automated systems are using selective and biased data to train their algorithms that have the potential to create new and reinforce historical gender and racial bias. The propagation of biases downstream—such as when DALL-E Mini and its equivalents are used in another application—can cause them to be entrenched and legitimized. To wit, the reification of these outputs can lead people to think their outputs are authoritative: one such example is when, say, DALL-E Mini and ChatGPT are used in tandem to author textbooks or other reference material. In the broader scheme of things, the distribution of gendered work—per labor statistics—are biased too, begging the bigger question: do we want AI systems to reflect our biased world or show us something that is more equal and aspirational?

Both technical and evaluative work in this field are emerging and urgent. Given the pace at which technologies and tools are being developed by Big Tech and unleashed on society, academic and ethical evaluation is always playing catch-up. Intense competition in the tech market incentivizes companies to release products and tech 'to market', as quickly as possible, removing any obstacles or processes that could slow down this process, including abandoning any beneficial processes in pursuit of markets. For instance, when this paper was first drafted, OpenAI could still lay some claim to its namesake. Earlier this year, Microsoft took a 49% stake in OpenAI and released ChatGPT and an integrated generative AI / search system with components from both GPT and Bing. Sadly, Microsoft has laid off its AI ethics team, due to pressure to get newer versions of AI models out to consumers quickly (53), as the ethics team was purportedly "slowing down innovation" (54).

These developments have been met with an ambivalent melange of wonder, derision, and apprehension. In the coming months and years, we are almost guaranteed to see further advancements in generative AI, as evident in the myriad of successors to DALL-E Mini (including DALL-E 2, Stable Diffusion, Midjourney, and its various derivatives), which far outpaces the existing speed at which rigorous ethical impact evaluations (such as this paper) could be feasibly produced.

In the meantime, we are concerned that virtually unregulated industry is increasingly taking a "ship first and ask questions later" approach to the software and models it releases to (or, pessimistically speaking, inflicts on) society. Tech companies are also prone to 'absolving themselves' from being accused of bias by blaming decisions on the 'machine' itself. Enforceable oversight by experts in computing, social sciences, and humanistic disciplines such as philosophy is clearly needed. In the United States and Europe, there have been moves in this direction, e.g., through the release of the Blueprint for an AI Bill of Rights by the Biden White House and related efforts by the European Commission. Given the potential for generative AI to reproduce and further entrench noxious social biases, these developments are necessary and urgent.

Limitations. We acknowledge several inherent limitations of the current work. First, we ensured that all the authors involved in coding the DALL-E Mini images come from a diverse range of backgrounds, disciplines, and life experiences, to minimize the risk of bias in coding the images. Nonetheless, we acknowledge that there is no surefire way of removing all human bias from the subjective coding process. Our current work is based on a binarized categorization when evaluating for gender- and racial-bias; however, in the spirit of (55), we understand that it is important to move beyond these binaries. Indeed, binary conceptions of gender and race in and of themselves embed various biases, contributing to the continued marginalization of those who don't easily fit within fixed categories. Further work includes looking at the intersectional factors surrounding stereotypes in image generation AIs, and expanding the corpora of seed words/phrases beyond occupational descriptors. In addition, several methods for debiasing datasets—predominantly for classification of structured data—do exist, but extant work for debiasing generative AIs are few and far between. Future work will look at efforts in this area, for example how DALL-E 2's online API approaches the issue of debiasing output.

Materials and Methods

At a high level of abstraction, our methodology consists of the following steps, in order:

1. Based on existing literature, producing a 'seed list' of phrases of terms, which represent occupations and job descriptions (e.g., doctor, teacher).
2. Feeding the 'seed list' into DALL-E Mini to generate 10 images per prompt.
3. Dividing the images amongst coders, who then code the images based on a unified codebook. Inter-coder agreement is measured, and the final result of coding is used as ground truth.
4. Determining, based on actual labor market and demographic statistics, whether the AI-generated images are representative of the demographics found in the real world.

Pre-registration. Before commencing the analysis proper, we pre-registered our hypotheses on the Open Science Framework (OSF) repository, at <https://osf.io/nft9p/registrations>.

Occupations and Prompt Generation. A novel approach to interrogating the bias found within a complex generative model is to determine how correlated a particular occupation or job description is with inherent societal biases.

Extant papers pave the way to our understanding of biases in computerized generative systems. As a result, we have identified a list of 105 occupations/job descriptors from similar studies dealing with gender or racial biases in image recognition and classification (40, 56) and text classification (39) systems. A paper on the subject (57) from a Science and Technology Studies (STS) perspective also provided us with similar bias-prone occupations. (The final list of 105 occupations is listed in ***SI Appendix***).

Image Generation. The creation of each image involved feeding various text prompts into our instance of DALL-E mini on a Google Colab Python notebook in the cloud. We refrained from using the ready-made, public-facing app (at craiyon.com) to avoid overloading the free service at cost to its creators. For reproducibility and to ensure faithfulness to the extant Craiyon app, we used the source code from the official DALL-E mini GitHub repository (2) (https://github.com/borisdayma/dalle-mini/blob/main/tools/inference/inference_pipeline.ipynb). All images were generated using the snapshot of code as of July 2022, specifically the parameters:
`DALLE_MODEL` = "`dalle-mini/dalle-mini/mega-1-fp16:v14`"
(commit "`9f723538131280eed9b96170176d95be`") and
`VQGAN_REPO` = "`dalle-mini/vqgan_imagenet_f16_16384`"
(commit "`e93a26e7707683d349bf5d5c41c5b0ef69b677a9`").

Coding and Evaluation. A total of 1,050 images were generated by requesting DALL-E for 10 images per prompt. The coder team, comprising a subset of this paper's authors, come from a variety of genders, ethnicities, age groups, and backgrounds, in order to reduce bias in the coding process.

Each image in each dataset was then coded by three separate coders, with subsets of images distributed randomly. A detailed example – of the instructions and images to code – is provided to coders is ***SI Appendix***.

To determine the reliability of these classifications, inter-rater reliability scores are calculated using Fleiss's multirater kappa in IBM SPSS Statistics.

ChatGPT. ChatGPT has been used to generate a clearly-indicated paragraph in the Introduction to illustrate its capabilities in context. See Footnote †.

ACKNOWLEDGMENTS. [Removed for review]

1. R Merritt, What is a transformer model? (NVidia Corporation) (2022).
2. B Dayma, et al., DALL·E mini (2021).
3. L Bouchard, How does dalle-mini work? (https://www.louisbouchard.ai/dalle-mini/) (2022) Accessed: 2022-7-28.
4. B Dayma, et al., DALL-E mini explained (https://wandb.ai/dalle-mini/dalle-mini/reports/DALL-E-Mini-Explained-with-Demo--Vmlldzo4NjIxODA) (2022) Accessed: 2022-7-28.
5. P Esser, R Rombach, B Ommer, Taming transformers for High-Resolution image synthesis. *arXiv* (2020).
6. A Lavelli, F Sebastiani, R Zanoli, Distributional term representations: an experimental comparison in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04. (Association for Computing Machinery, New York, NY, USA), pp. 615–624 (2004).
7. SU Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism.* (NYU Press), (2018).
8. C O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy.* (Broadway Books), (2016).
9. Australian Human Rights Commission, Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias (2020), (Australian Human Rights Commission), Technical report (2020).
10. BD Mittelstadt, P Allo, M Taddeo, S Wachter, L Floridi, The ethics of algorithms: Mapping the debate. *Big Data & Soc.* **3**, 2053951716679679 (2016).
11. A Tsamados, et al., The ethics of algorithms: key problems and solutions. *AI Soc.* (2021).
12. SY Liao, B Huebner, Oppressive things. *Philos. Phenomenol. Res.* **103**, 92–113 (2021).
13. S Buranyi, Rise of the racist robots – how AI is learning all our worst impulses. *The Guard.* (2017).
14. Z Obermeyer, B Powers, C Vogeli, S Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
15. T Panch, H Mattie, R Atun, Artificial intelligence and algorithmic bias: implications for health systems. *J. Glob. Heal.* **9**, 010318 (2019).
16. K Martin, Ethical implications and accountability of algorithms. *J. Bus. Ethics* **160**, 835–850 (2019).
17. MV Santelices, M Wilson, Unfair treatment? the case of freedle, the SAT, and the standardization approach to differential item functioning. *Harv. Educ. Rev.* **80**, 106–134 (2010).
18. I Ajunwa, SA Friedler, C Scheidegger, S Venkatasubramanian, Hiring by algorithm: predicting and preventing disparate impact. (2016).
19. M Raghavan, S Barocas, J Kleinberg, K Levy, Mitigating bias in algorithmic hiring: evaluating claims and practices in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20. (Association for Computing Machinery, New York, NY, USA), pp. 469–481 (2020).

20. J Larson, J Angwin, T Parris, Jr, Breaking the black box: How machines learn to be racist (https://www.propublica.org/article/breaking-the-black-box-how-machines-learn-to-be-racist) (2016) Accessed: 2022-8-31.
21. R Wexler, Code of silence (https://washingtonmonthly.com/2017/06/11/code-of-silence/) (2017) Accessed: 2022-8-31.
22. LN Guo, MS Lee, B Kassamali, C Mita, VE Nambudiri, Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-a scoping review. *J. Am. Acad. Dermatol.* **87**, 157–159 (2022).
23. MS Lee, LN Guo, VE Nambudiri, Towards gender equity in artificial intelligence and machine learning applications in dermatology. *J. Am. Med. Inform. Assoc.* **29**, 400–403 (2022).
24. O Kharif, No credit history? no problem. lenders are looking at your phone data. *Bloom. News* (2016).
25. A Howard, J Borenstein, The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Sci. Eng. Ethics* **24**, 1521–1536 (2018).
26. J Dastin, Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018).
27. M Langenkamp, A Costa, C Cheung, Hiring fairly in the age of algorithms. (2019).
28. L Alexander, Do google's 'unprofessional hair' results show it is racist? *Guardian* (2016).
29. BBC News, Google apologises for photos app's racist blunder. *BBC* (2015).
30. A Hern, Google's solution to accidental algorithmic racism: ban gorillas (https://amp.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people) (2018) Accessed: 2022-7-28.
31. L Sweeney, Discrimination in online ad delivery. *ACM Queue* **11**, 10–29 (2013).
32. J Dressel, H Farid, The accuracy, fairness, and limits of predicting recidivism. *Sci Adv* **4**, eaao5580 (2018).
33. C Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
34. PM Asaro, AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technol. Soc. Mag.* **38**, 40–53 (2019).
35. S Changpinyo, P Sharma, N Ding, R Soricut, Conceptual 12m: Pushing Web-Scale Image-Text Pre-Training to recognize Long-Tail visual concepts. *arXiv* **cs.CV** (2021).
36. M Mitchell, et al., Model cards for model reporting in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. (Association for Computing Machinery, New York, NY, USA), pp. 220–229 (2019).
37. B Dayma, et al., DALL·E mini model card (https://huggingface.co/dalle-mini/dalle-mini) (2022) Accessed: 2022-7-28.
38. M Vlasceanu, DM Amodio, Propagation of societal gender inequality by internet search algorithms. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2204529119 (2022).
39. M De-Arteaga, et al., Bias in bios: A case study of semantic representation bias in a High-Stakes setting in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. (Association for Computing Machinery, New York, NY, USA), pp. 120–128 (2019).
40. J Cho, A Zala, M Bansal, DALL-Eval: Probing the reasoning skills and social biases of Text-to-Image generative transformers. *arXiv* **cs.CV** (2022).
41. R Steed, A Caliskan, Image representations learned with unsupervised Pre-Training contain human-like biases in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. (Association for Computing Machinery, New York, NY, USA), pp. 701–713 (2021).
42. A Birhane, Algorithmic injustice: a relational ethics approach. *Patterns* **2**, 100205 (2021).
43. UB of Labor Statistics, Employment by detailed occupation, 2019 and projected 2029 (https://www.bls.gov/cps/cpsaat11.htm) (2020) Accessed: Mar. 8, 2023.
44. S Sczesny, M Formanowicz, F Moser, Can Gender-Fair language reduce gender stereotyping and discrimination? *Front. Psychol.* **7**, 25 (2016).
45. M Bogen, All the ways hiring algorithms can introduce bias. *Harv. Bus. Rev.* (2019).
46. P Hegarty, C Buechel, Androcentric reporting of gender differences in APA journals: 1965–2004 (2006).
47. S Njoto, Gendered bots? bias in the use of artificial intelligence in recruitment, (The Policy Lab, The University of Melbourne), Technical report (2020).
48. S Njoto, et al., Gender bias in AI recruitment systems: A sociological- and data science-based case study in *Proceedings of the 2022 IEEE International Symposium on Technology and Society (ISTAS).* (2022).
49. C Basta, MR Costa-jussà, N Casas, Evaluating the underlying gender bias in contextualized word embeddings in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing.* (Association for Computational Linguistics, Stroudsburg, PA, USA), (2019).
50. T Manzini, L Yao Chong, AW Black, Y Tsvetkov, Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* (Association for Computational Linguistics, Minneapolis, Minnesota), pp. 615–621 (2019).
51. T Bolukbasi, KW Chang, J Zou, V Saligrama, A Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* (year?).
52. YC Tan, L Elisa Celis, Assessing social and intersectional biases in contextualized word representations. *arXiv* (2019).
53. R Bellan, Microsoft lays off an ethical AI team as it doubles down on OpenAI. *TechCrunch* (2023).
54. E Ajao, Reasons for and effects of microsoft cutting AI ethics unit (https://www.techtarget.com/searchenterpriseai/news/365532615/Reasons-for-and-effects-of-Microsoft-cutting-AI-ethics-unit) (2023) Accessed: 2023-3-20.
55. W Guo, A Caliskan, Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *arXiv* (2020).
56. C Schwemmer, et al., Diagnosing gender bias in image recognition systems. *Socius* **6**, 2378023120967171 (2020).
57. K Crawford, T Paglen, Excavating AI: The politics of images in machine learning training sets (https://excavating.ai/) (year?) Accessed: 2023-2-6.