A Friendly Face: Do Text-to-Image Systems Rely on Stereotypes when the Input is Under-Specified?

Kathleen C. Fraser, Svetlana Kiritchenko, and Isar Nejadgholi

National Research Council Canada
Ottawa, Canada
{kathleen.fraser, svetlana.kiritchenko, isar.nejadgholi}@nrc-cnrc.gc.ca

Abstract

As text-to-image systems continue to grow in popularity with the general public, questions have arisen about bias and diversity in the generated images. Here, we investigate properties of images generated in response to prompts which are visually under-specified, but contain salient social attributes (e.g., 'a portrait of a threatening person' versus 'a portrait of a friendly person'). Grounding our work in social cognition theory, we find that in many cases, images contain similar demographic biases to those reported in the stereotype literature. However, trends are inconsistent across different models and further investigation is warranted.

Introduction

Recent advances in natural language processing and computer vision have led to the development of text-to-image systems with unprecedented levels of realism and flexibility. At the same time, commentators have noted potential ethical issues related to the use of copyrighted artworks in the training sets, the generation of hateful and offensive content, as well as issues of bias and diversity in the model outputs. Relating to the latter, research work has begun to audit the output of such models, investigating stereotypical associations between occupations and particular races and genders (Cho, Zala, and Bansal 2022), as well as between the word "American" and lighter skin colours (Wolfe and Caliskan 2022).

Here, we take an alternative approach inspired by stereotype research in social psychology and focus on *perceived traits* of individuals. However, we approach the problem from the inverse direction of most psychological studies. Rather than treating a demographic group (say, women) as the independent variable, and asking respondents for the associated traits (say, nurturing and emotional), here we use the *trait* as a prompt to the text-to-image system, and observe the demographic properties of the resulting image. So, to continue our example: if we ask the system for an image of *an emotional person*, will it return mostly pictures of women?

We ground our investigation in the ABC Model of social cognition (Koch et al. 2016). This model proposes three basic dimensions of social judgement; namely: Agency (A), Beliefs (B), and Communion (C). These three dimensions can be further broken down into 16 polar traits. For example, Agency comprises traits such as *powerful vs. powerless* and *high-status vs. low-status*, Beliefs include traits such as *conservative vs. liberal* and *religious vs. science-oriented*, and Communion includes *sincere vs. dishonest* and *altruistic vs. egoistic*. This model suggests that all our stereotypes of different groups can be specified in this 3-dimensional space: e.g., in the North American context, Southerners may be stereotyped as laid-back, friendly, and religious (low agency, high communion, low beliefs[1]), while tech entrepreneurs may be stereotyped as wealthy, science-oriented, and greedy (high agency, high beliefs, low communion).

Clearly, these adjectives are under-specified with respect to a visual representation: what does a *powerful* person look like? What does a *sincere* person look like? It is precisely this under-specificity that can result in biased outputs, as the model must "fill in the blanks" with whatever cultural knowledge it has learned from the training data. However, as Hutchinson, Baldridge, and Prabhakaran (2022) point out, "Descriptions and depictions necessarily convey incomplete information about all but the most trivial scene." The model's approach to handling under-specification will therefore have varied and wide-ranging effects.

Thus, our research question is as follows: If we prompt the model to generate a person with particular social traits (as defined by the ABC Model), will the resulting images show the stereotypical demographic characteristics associated with those traits? We investigate this question using the 16 traits of the ABC Model and the demographic characteristics of skin colour, gender, and age, with three popular text-to-image models: DALL-E 2, Midjourney, and Stable Diffusion. We find that while not all traits generate stereotypical images, each model shows idiosyncratic biases along certain dimensions. We also observe intersectional biases, in particular a bias in all three systems associating the adjective of "poor" with darker-skinned males.

[1]While Agency and Communion have clear positive (high) and negative (low) poles, the Beliefs dimension is defined along a continuum from *progressive* to *conservative*, with *progressive* being arbitrarily assigned the "positive" direction. Note also that polarity does not necessarily align with normative judgements of *good/bad* behaviour; e.g., dominating people have *positive* agency, although their dominating behaviour would not necessarily be seen as *good*.

Related Work

In recent years, the machine learning research community has devoted significant effort to combating bias-related issues in computer vision applications. One line of work has analyzed the biases that stem from unbalanced class distributions in training datasets, resulting in systematic errors and poor performance on the minority classes. For example, as Buolamwini and Gebru (2018) reported, the over-representation of light-skinned individuals in commonly-used facial recognition datasets leads to drastically larger error rates of commercial gender classification systems for darker-skinned females. Data augmentation methods can improve data balance, the efficacy of which is often measured by overall accuracy as well as more uniform performance across attributes such as race and gender (Deviyani 2022). In a recent work, Mitchell et al. (2020) introduced quantitative metrics to directly measure the diversity and inclusion of a dataset, defining these concepts with respect to sociopolitical power differentials (gender, race, etc.) in management and organization sciences. Other works studied biases originated from annotations, such as linguistic biases, stereotypical descriptions, and unwarranted inferences about people's demographic traits in crowd-sourced annotations (van Miltenburg 2016), or reported bias when annotators make implicit decisions about what is worth mentioning in the annotations (Misra et al. 2016).

Besides data and annotation distributions, the choice of models can impact the fairness of trained algorithms. For example, computer vision models trained with zero-shot natural language supervision exhibit unexpected systematic errors associated with gender, race, and age traits, for specific design choices (Agarwal et al. 2021). Also, image captioning models that learn to use contextual cues often exaggerate stereotypical cues present in the context to predict demographic traits (Hendricks et al. 2018). These observations call for task-specific and safety-focused evaluations to audit computer vision models for biased outcomes before deployment. Raji et al. (2020) identified multiple ethical concerns in auditing commercial face recognition systems and recommended deliberate fairness evaluations as minimizing biases for some groups might cause unintended harms for others.

Recently, work has begun that focuses on particular biases that have emerged in multi-modal language–vision machine learning systems. Wolfe and Caliskan (2022) reported that racial biases about American identity, previously observed in social psychology, are learned by multi-modal embedding models and propagated to downstream tasks. Other works proposed evaluation frameworks to assess biases in text-to-image systems (Cho, Zala, and Bansal 2022) or training mechanisms such as adversarial learning to reduce representation biases in language–vision models (Berg et al. 2022). Further, in a position paper, Hutchinson, Baldridge, and Prabhakaran (2022) discussed social bias amplification, among other ethical concerns that arise from the use of text-to-image systems. They identified ambiguity and under-specification as the root causes of these risks and proposed conceptual frameworks to manage them. Specifically, they introduced two approaches to deal with under-specification: *Ambiguity In, Ambiguity Out* (AIAO) and *Ambiguity In, Diversity Out* (AIDO). In the AIAO approach, the model is encouraged to generate ambiguous images when the concepts in input text are under-specified. In the alternative approach, AIDO, the preferable behaviour is generating a set of maximally diverse images to cover the space of possibilities for the under-specified concept. We consider both of these approaches in our current analyses.

Social Stereotypes

We draw our hypotheses from the existing survey-based literature on prevalent stereotypes in North American society. In the paper introducing the ABC model, Koch et al. (2016) present results that place various social groups in the three-dimensional Agency-Beliefs-Communion space. Most relevant to our work here are social groups defined by gender, age, or skin colour. While the specific results vary somewhat across their sub-studies, some consistent patterns are seen: in terms of Agency, white people are rated higher than people of colour, old or elderly people are rated higher than young people, and men are rated higher than women. In terms of Beliefs, young people are rated as more progressive than older people, and there is no obvious distinction based on gender or skin colour. In terms of Communion, white people are rated higher than Black people, although Latinos, Hispanics, Asians, and Indians are also rated as positive-communion. Older people are also seen as higher on this dimension than younger people. The experimental design does not directly compare communion values for men and women. However, other related literature confirms many of these predictions and also reports that women are seen as more Communal (warm, friendly) than men (Fiske et al. 2002; Nicolas, Bai, and Fiske 2022), and white people are seen as more modern and science-oriented (high-Beliefs) than Black people (Cao et al. 2022).

As a result, our hypotheses are as follows. Given the under-specification of our prompts with respect to the demographic characteristics of the generated image subject, the text-to-image models will default to social stereotypes learned from the data, namely:

- **High-agency** words will tend to generate images of people with lighter skin, older age, and male gender, while **low-agency** words will tend to generate images of people with darker skin, younger age, and female gender.
- **High-belief** (progressive) words will tend to generate images of younger and lighter-skinned people, while **low-belief** (conservative) words will tend to generate images of older and darker-skinned people.
- **High-communion** words will tend to generate images of people with lighter skin, older age, and female gender, while **low-communion** words will tend to generate images of people with darker skin, younger age, and male gender.

Methodology

We first describe the three contemporary text-to-image systems evaluated in the current study, and then provide details on image generation and annotation processes.

Agency	powerless ↔ powerful low-status ↔ high-status dominated ↔ dominating poor ↔ wealthy meek ↔ confident passive ↔ competitive	**Belief**	religious ↔ science-oriented conventional ↔ alternative conservative ↔ liberal traditional ↔ modern	**Communion**	untrustworthy ↔ trustworthy dishonest ↔ sincere unfriendly ↔ friendly threatening ↔ benevolent unpleasant ↔ likable egoistic ↔ altruistic

Table 1: List of stereotype dimensions and corresponding traits in the ABC model, adapted from Cao et al. (2022).

Text-to-Image Systems

All three systems evaluated in this study, DALL-E 2,[2] Midjourney,[3] and Stable Duffusion,[4] generate original images from textual prompts and/or uploaded images. They are based on state-of-the-art image generation technology, like diffusion models (Sohl-Dickstein et al. 2015; Nichol et al. 2022) and CLIP image embeddings (Radford et al. 2021), and are trained on millions and billions of text–image examples scraped from the web. We briefly discuss each system and provide more details in the Appendix. All images used in the study were generated in October 2022.

DALL-E 2: This is a research and production system released as beta version by OpenAI in July 2022. DALL-E 2 (hereafter, simply 'DALL-E') aims to create photorealistic, diverse images, that closely represent the textual prompts (Ramesh et al. 2022). To more accurately reflect the diversity of the world's population and to prevent the dissemination of harmful stereotypes, the system has been extended to further diversify its output for under-specified prompts of portraying a person (e.g., 'a portrait of a teacher').[5]

Midjourney: This system was created by an independent research lab Midjourney and released as beta version in July 2022; we used the most recent version, v3. It has been designed as a social app where users generate images along side other users in public community channels through the chat service Discord. The system details have not been publicly released.

Stable Diffusion: This system was publicly released by Stability AI under a Creative ML OpenRAIL-M license in August 2022. It is based on latent diffusion model by Rombach et al. (2022). The system was trained to produce aesthetically pleasing images using LAION-Aesthetics dataset. We accessed Stable Diffusion v1.5 through the DreamStudio API with default settings.

Image Generation

For each of the three systems, we used templates to produce prompts containing each of the adjectives in Table 1. These adjectives were taken from Koch et al. (2016) with a few minor variations: the original ABC Model uses the adjectives *warm*, *cold*, and *repellent*, which we found to be too semantically ambiguous to produce reliable results (e.g., would generate a person who was physically freezing cold). These words were replaced with *friendly*, *unfriendly*, and *unpleasant*, respectively. Koch et al. (2016) also use two words which have extremely low frequency (less than one instance per million in the SUBTLEX-US corpus), namely *unconfident* and *unassertive*; these were replaced with *meek* and *passive*.

Our basic prompt took the form of: `portrait of a ⟨adjective⟩ person`. The word `portrait` cues the system to show the face of the subject, which contains most of the visual information needed to make the demographic annotations. For each model, we found that we needed to adapt the prompt slightly to achieve acceptable results. For DALL-E, the prompt as written was very effective in generating colour, photo-realistic results; in the few cases that a drawing or black-and-white image were generated, we reran the generation until a colour, photo-realistic result was achieved. For Midjourney, the basic prompt had a tendency to generate more artistic interpretations. Adding the flag `--testp` helped generate photo-realistic results, but they were mostly black-and-white. Adding the keywords `color photograph` was not effective, as it generated highly stylized colours that obscured the skin colour of the subjects. Instead, we found the keywords `Kodak Portra 400` (a popular colour film stock) to be highly effective at producing colour, photorealistic results. Similarly for Stable Diffusion, adding the `Kodak Portra 400` keywords to the end of the prompt led to the generation of interpretable results as opposed to painterly, abstract images.

Since DALL-E outputs images in batches of 4, we decided to generate 24 images per trait (12 per pole of each trait). We additionally generated 24 baseline images for each model, using the basic prompt of `portrait of a person`, with no trait adjective. Thus for each of the three models, we generated a total of 408 images.

Image Annotation

Each image was annotated by three annotators (the authors of the paper). Our demographic characteristics of interest were gender, skin colour, and age. The process of inferring demographic characteristics from images has numerous ethical challenges. We outline our processes and assumptions here, with a more detailed discussion in the Appendix.

First, we emphasize that we are annotating perceived demographic characteristics of *AI-generated images*, not real people. We are doing this with the goal of assessing the diversity of the outputs, not of categorizing real individuals. To that end, we have also decided not to make our annotations

[2]https://openai.com/dall-e-2/

[3]https://www.midjourney.com

[4]https://huggingface.co/spaces/stabilityai/stable-diffusion

[5]https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/

Dataset	Gender	Skin Colour	Age
Midjourney	0.84	0.57	0.76
DALL-E	0.92	0.64	0.68
Stable Diffusion	0.75	0.54	0.54

Table 2: Cohen's Kappa (κ) metric of inter-annotator agreement for each demographic variable and each dataset.

Dataset	Gender	Skin Colour	Age
Midjourney	0.06	0.06	0.08
DALL-E	0.00	0.03	0.04
Stable Diffusion	0.06	0.01	0.11

Table 3: Proportion of images for which at least one annotator indicated that no visual cues to the given demographic variable were present in the image, reflecting an 'Ambiguity In, Ambiguity Out' (AIAO) strategy.

publicly available, so as not to facilitate such a use case.

Second, following best practices from the literature (Buolamwini and Gebru 2018), we do not attempt to categorize particular races/ethnicities, but rather focus on the more objective measure of skin colour (from "lighter" to "darker"). We also recognize gender as being a non-binary variable and allow for a gender-neutral annotation.

Finally, we combine the annotations using an averaging technique such that each annotator's judgement is equally weighted, rather than using a majority-voting scheme. The full annotation instructions are available in the Appendix. Briefly, each demographic variable can receive one of four possible annotations (gender: male, female, gender neutral, or no gender information available; skin colour: darker, lighter, in-between, or no skin colour information available; age: older, younger, in-between, or no age information available). These categorical annotations are converted to numerical values and averaged over the three annotators. As a concrete example, if two annotators marked an image subject as having darker skin (+1) and one annotated it as in-between (0), then the image would be assigned a summary skin colour value of 0.67.

Results

Annotation Reliability

The Cohen's Kappa values for inter-annotator agreement are given in Table 2. In general, the gender annotation had highest agreement, and the skin colour annotation had the lowest. This can be partly attributed to the fact that perceived gender typically fell into either 'male' or 'female' categories, with only a few gender-ambiguous images in between, while skin colour ranged across a full spectrum, creating more annotation uncertainty between categories. (Although, note that this issue is partially alleviated by our annotation averaging technique.) Furthermore, skin colour estimation was confounded by varying lighting conditions in the images. The agreement values are overall lower for the Stable Diffusion dataset, reflecting a qualitatively lower degree of photorealism in the generated images.

Ambiguity In, Ambiguity Out

As mentioned above, one viable strategy for dealing with ambiguous or under-specified inputs is to in turn produce ambiguous or under-specified outputs (AIAO). Our experimental paradigm constrained the systems' ability to deploy this strategy, by first prompting for a portrait (implying that the face should be visible), by constraining the analysis to colour images, and by prompting for photographic-style results. Loosening these constraints would no doubt

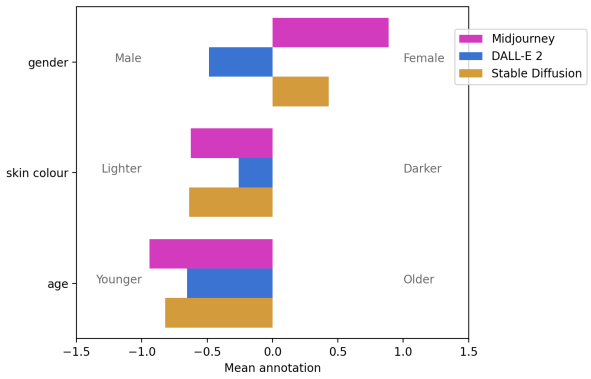


Figure 1: Baseline values for gender, skin colour, and age for the three models.

lead to more creative and interpretative outputs by the models. Nonetheless, in some cases the generated images were ambiguous with respect to one or more of the demographic variables. Common methods of creating ambiguity included: positioning the subject facing away from the camera, obscuring the subject's face with an object, generating non-realistic skin colours (e.g., purple), or blurring the subject's features. The rate at which each model produced such AIAO images is given in Table 3. In the following sections, we remove these images from the analysis and focus on the alternate strategy of Ambiguity In, Diversity Out (AIDO).

Baseline Results

Figure 1 shows the baseline results of prompting each model for `portrait of a person`. Midjourney and Stable Diffusion have a tendency to produce female subjects, while DALL-E tends to produce males. All three systems have a tendency to generate subjects with lighter skin tones, although DALL-E less so than the other two. All three systems also have a strong tendency to produce younger-looking subjects.

ABC Model Results

Figure 2 shows, for each demographic variable, the average annotation across all images associated with the positive and negative poles in each ABC dimension. That is, for Agency, the positive pole includes all images generated from *powerful*, *dominating*, *high-status*, etc., and the negative pole includes all images generated from *powerless*, *dominated*,

low-status, etc., and similarly for Beliefs and Communion. Thus for Agency, each of the positive and negative values are averaged over 3 annotators per image $\times$ 12 images per trait $\times$ 6 traits, or 216 annotations. When there is a significant difference between the positive and negative poles, this is indicated with an asterisk in the figure ($p < 0.05$, Mann-Whitney U-Test). This comparison corresponds to the concept of stereotype ***direction*** as defined by Nicolas, Bai, and Fiske (2022): "Group X is associated with positive or negative direction on a given trait." However, that paper also introduces the complementary notion of ***representativeness***, i.e., Group X being highly associated with a given trait, regardless of polarity. We discuss the results with reference to both of these concepts in the following.

Beginning with the gender variable, in Fig. 2a we observe a significant difference in the Midjourney results for Agency, with high-agency words more likely to generate images representing male gender, and low-agency words more likely to generate images representing female gender, as hypothesized. There is also a significant difference in Communion, with low-communion words more associated with male gender as expected. When we break these dimensions down by trait (shown in the Appendix), this corresponds to more images of men generated for words like *high-status* and *dominating* (high-Agency) and *dishonest* and *unpleasant* (low-Communion), while more images of women were generated for adjectives like *powerless*, *friendly*, and *likable*. In the case of DALL-E (Fig. 2b), there is no significant difference on gender along any of the three dimensions. However, we do observe a difference in *representativeness*: specifically, that DALL-E has a tendency to produce more males than females for all ABC dimensions, regardless of polarity. For Stable Diffusion (Fig. 2c), similar to Midjourney we observe that low-communion words are significantly associated with male gender, and similar to DALL-E that the male gender is over-represented in general.

Turning now to the variable of skin colour, for Midjourney (Fig. 2d) we observe a significant difference along the Beliefs dimension, with progressive beliefs more highly-associated with lighter skin colours, as expected. This trend is driven by a high proportion of lighter-skinned subjects generated for the prompts *science-oriented* (versus *religious*) and *modern* (versus *traditional*). In terms of representativeness, all dimensions have a tendency towards lighter skin. For DALL-E (Fig. 2e), for Agency and Beliefs we see no significant difference in direction with respect to skin colour, and more equal representation. However, there is a significant difference in direction for Communion, with low-communion words more associated with lighter skin, in contradiction to our hypothesis. This trend is driven by adjectives *untrustworthy*, *threatening*, and *unpleasant*. In the case of Stable Diffusion (Fig. 2f), we again observe an over-representation of images depicting lighter-skinned subjects in all dimensions. Additionally, the system shows a significant difference in the dimensions of Beliefs (progressive beliefs more associated with lighter skin) and Communion (low communion more associated with lighter skin, specifically for the words *dishonest* and *threatening*).

Finally, considering age: Midjourney (Fig. 2g) shows a significant difference in Agency and Communion, with low-agency words and high-communion words more associated with images of younger people (recall, the first trend is in keeping with our hypotheses but the second is not). For DALL-E (Fig. 2h), the positive poles of all three ABC dimensions are associated with younger age, and this difference is significant in all cases. For Stable Diffusion (Fig 2i), the same trend occurs, although it is only significant in the Beliefs dimension. However, for Stable Diffusion we also note a highly-skewed representativeness towards younger age in all dimensions. In particular, for all three systems, the adjective *likable* was highly associated with younger age, with its contrast adjective *unpleasant* ranging from moderately to highly associated with older age.

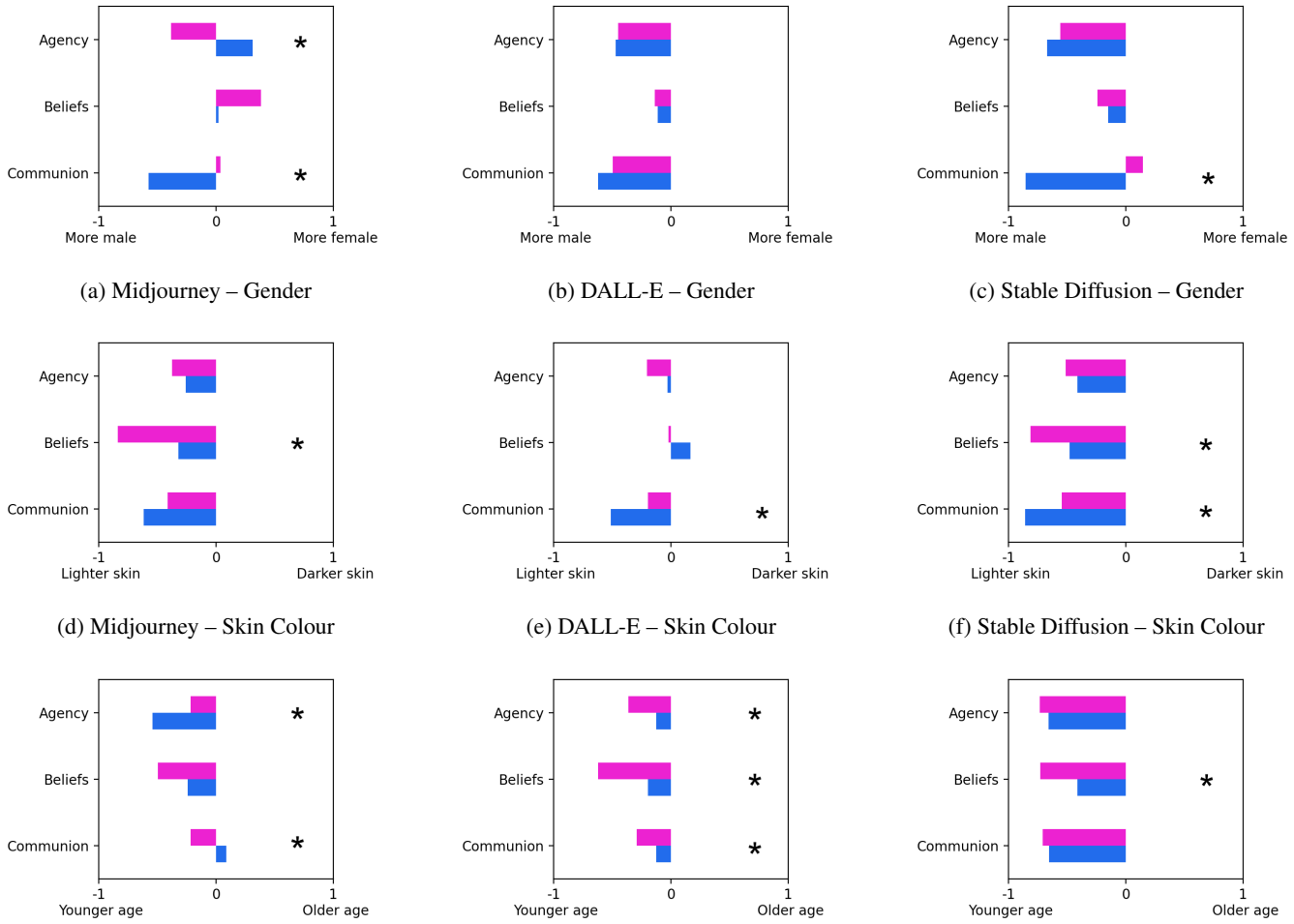
Intersectional Results

While looking at each dimension individually is informative, additional insight can be obtained by considering the results intersectionally. Figure 3 shows the average skin colour and gender annotation for each pole of each trait. In Figure 3a it is obvious that while Midjourney generates images across the range of genders, the only case where the model had a strong tendency to generate males with darker skin was for the adjective *poor*. It had a slight tendency to generate darker-skinned males for *competitive* (with images typically showing athletes), and a slight tendency to generate darker-skinned women for *traditional* (with images showing women in "traditional" dress from various cultures).

For DALL-E (Fig. 3b), we observe a different distribution. DALL-E is more likely to generate both darker- and lighter-skinned people, but those people tend to be male. Images of lighter-skinned women were generated for the adjective *alternative*, and images of darker-skinned women were generated for *traditional*, as above. While the distribution across skin colour is more equitable in general, we do again note that the adjective *poor* tends to generate darker-skinned males.

Turning to the Stable Diffusion results in Figure 3c, we see some similarities and some differences in comparison with the other models. Once again, darker skinned males are generated mostly by the adjective *poor*, with a slight trend for *competitive*, and again darker-skinned females are associated with *traditional*. Unlike DALL-E, there is a higher density of points in the lighter-skinned female quadrant, and unlike Midjourney the points all tend to be associated with 'positive' adjectives: *benevolent* and *sincere* occur right along the 0 axis for skin colour, with *powerful* and *likable* associated with lighter-skinned women.

For lack of space, the corresponding figures plotting age versus skin colour and age versus gender are given in the Appendix, but we briefly summarize the findings here. Figure A.4 shows that the adjective *poor* is also anomalous when we consider age versus skin colour: in the case of both DALL-E and Stable Diffusion, it is the point closest to the upper right (oldest and darkest-skinned). Overall, Midjourney outputs a range of ages, but primarily lighter skin colours (points concentrated on the left-most quadrants), DALL-E produces a range of skin colours but mostly younger faces (points concentrated in the bottom two



(a) Midjourney – Gender (b) DALL-E – Gender (c) Stable Diffusion – Gender

(d) Midjourney – Skin Colour (e) DALL-E – Skin Colour (f) Stable Diffusion – Skin Colour

(g) Midjourney – Age (h) DALL-E – Age (i) Stable Diffusion – Age

Figure 2: Average annotation values for gender, skin colour, and age, for each of the ABC dimensions and for each model. The positive pole of each trait is shown in pink; the negative pole is shown in blue. Significant differences between positive and negative traits are indicated with an asterisk.

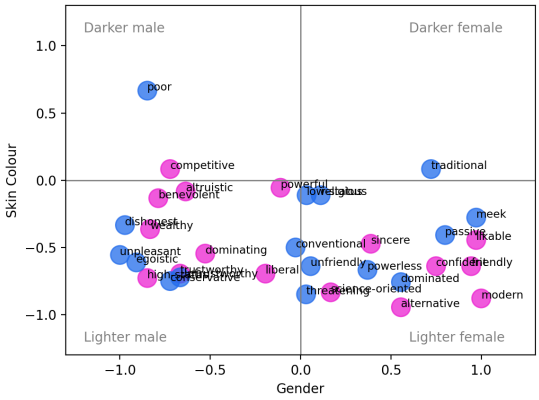
quadrants), and Stable Diffusion produces mostly lighter, younger faces (points mostly in the bottom, left quadrant).

When we consider age versus gender (Fig. A.5), Midjourney shows a surprisingly linear negative trend, with some adjectives associated with older males, and others associated with younger females, but no traits associated primarily with older females, and only one trait (*competitive*) associated primarily with younger males. DALL-E and Stable Diffusion both exhibit a trend of generating younger males.

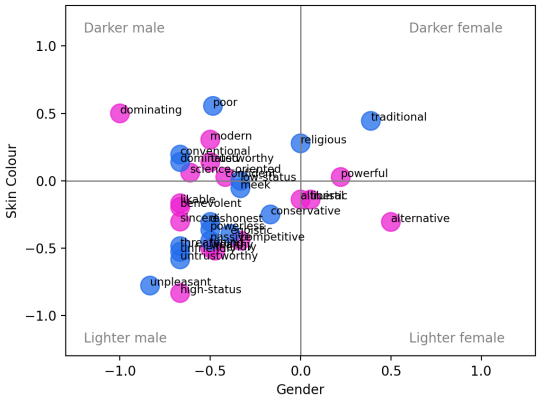
Summary and Discussion

Many of the significant differences in Figure 2 did confirm our hypotheses: high-Agency words generated more men than women, as did low-Communion words, and in the case of Midjourney, low-Agency words were associated with younger age. As well, both Midjourney and Stable Diffusion showed a significant tendency to associate progressive Beliefs with lighter skin, primarily driven by the traits *modern–traditional* and *science-oriented–religious*, as also reported by (Cao et al. 2022). In contrast to our hypotheses, lighter skin was associated with low-communion adjectives for both DALL-E and Stable Diffusion. We also found an unexpected trend of high-communion associated with younger age. Combined with the fact that all three models showed a preference for generating more images of younger people, it appears that age-based bias may need to be addressed by the developers of such systems.

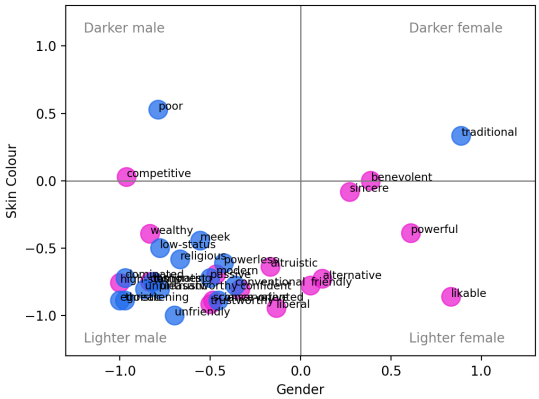
Considering the intersectional results, the lack of representation of darker females is particularly striking, and unfortunately not surprising, given previous scholarship on intersectional bias in AI systems (D'ignazio and Klein 2020). That all three systems also associated poverty with dark-skinned males warrants further investigation. Examples of the images produced for prompts *poor* and *wealthy* are given



(a) Midjourney



(b) DALL-E



(c) Stable Diffusion

Figure 3: Intersectional view of each trait (skin colour x gender). The positive poles of each trait are shown in pink; the negative poles are shown in blue.

in the Appendix, Figure A.6.

These results must be considered preliminary, due to the small sample size and limited variety in the prompt structure. However, we offer a brief discussion on some of the factors that may contribute to the differences seen across the three models, as well as differences with the ABC Model.

First, we must consider the effect of how the training data was collected: largely by scraping the web for image-caption pairs. Hutchinson, Baldridge, and Prabhakaran (2022) describe the many different relationships between captions and images that can exist, ranging from simple descriptions of the visual content of the image, to providing interpretations, explanations, or additional information. Misra et al. (2016) discuss the human reporting bias that exists in image datasets, as annotators choose what is "worth mentioning" in an image. In particular, traits which are seen as stereotype-consistent or "default" are often not mentioned; for example, van Miltenburg (2016) reports that the race of babies in the Flickr30k image dataset is not typically mentioned, unless the baby is Black or Asian. White babies are assumed to be the default. Hence, some of the stereotypical associations that we do *not* see may potentially be because internet-users do not explicitly mention default traits in image captions.

Another factor is the intended use case for these text-to-image systems. Midjourney and Stable Diffusion in particular have been marketed as methods of generating AI artwork. Therefore, the systems are optimized to produce "aesthetically-pleasing" results, with users providing feedback on what exactly that means to them. Historically, Western art has been biased towards images of the female form (Nead 2002), as well as aesthetic preferences for youth and beauty. Therefore it is also possible that some of this aesthetic bias for what makes "beautiful" art is affecting the output distribution of these systems.

Finally, it should be acknowledged that the human creators of such systems make normative design choices. OpenAI has published blog posts describing their efforts to reduce bias and improve safety in DALL-E 2, and in Figure 2 we do see fewer extreme disparities in representation, particularly with respect to skin colour. On the other hand, expressing a philosophy of user freedom, Stable Diffusion has rejected the approach of filtering input and/or output content, and puts the responsibility on the user to use the system appropriately. The tension between freedom of expression and harm reduction has been seen in many areas of artificial intelligence, and continues to be an open – and potentially unsolvable – question.

Conclusion and Future Work

Although not all systems showed the same types of stereotypical biases, each one demonstrated some room for improvement. In particular, we believe that analyzing age-related bias will be one fruitful area of research. This work also points to the need for further investigation of the relationships between race, gender, and economic status that have been encoded in such systems. Future work should involve confirming (or dis-confirming) the presence of such biases with bigger sample sizes and more varied prompt structure and content.

One key open question is to what extent the bias originates from the training data, the model architecture, or the model parameters. The answer to that question will help inform appropriate de-biasing methods at training time. Another promising avenue of research involves mitigating bias at inference time through careful prompt engineering. For example, Bansal et al. (2022) report that modifying prompts with phrases such as 'irrespective of gender' can encourage text-to-image models to generate outputs of various genders. The further development of such intervention strategies will also help improve the fairness and diversity of model outputs.

References

Agarwal, S.; Krueger, G.; Clark, J.; Radford, A.; Kim, J. W.; and Brundage, M. 2021. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

Bansal, H.; Yin, D.; Monajatipoor, M.; and Chang, K.-W. 2022. How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? *arXiv e-prints*, arXiv–2210.

Berg, H.; Hall, S. M.; Bhalgat, Y.; Yang, W.; Kirk, H. R.; Shtedritski, A.; and Bain, M. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proc. of AACL*.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. of the Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.

Cao, Y. T.; Sotnikova, A.; Daumé III, H.; Rudinger, R.; and Zou, L. 2022. Theory-grounded measurement of US social stereotypes in English language models. In *Proc. of NAACL*, 1276—-1295.

Cho, J.; Zala, A.; and Bansal, M. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*.

Deviyani, A. 2022. Assessing Dataset Bias in Computer Vision. *arXiv preprint arXiv:2205.01811*.

D'ignazio, C.; and Klein, L. F. 2020. *Data Feminism*. MIT Press.

Fiske, S. T.; Cuddy, A. J.; Glick, P.; and Xu, J. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6): 878—-902.

Hamidi, F.; Scheuerman, M.; and Branham, S. 2018. Can Gender Be Computed? *The Conversation US*.

Hanley, M.; Barocas, S.; Levy, K.; Azenkot, S.; and Nissenbaum, H. 2021. Computer vision and conflicting values: Describing people with automated alt text. In *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, 543–554.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 771–787.

Hutchinson, B.; Baldridge, J.; and Prabhakaran, V. 2022. Underspecification in Scene Description-to-Depiction Tasks. *arXiv preprint arXiv:2210.05815*.

Koch, A.; Imhoff, R.; Dotsch, R.; Unkelbach, C.; and Alves, H. 2016. The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110(5): 675.

Misra, I.; Lawrence Zitnick, C.; Mitchell, M.; and Girshick, R. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proc. CVPR*, 2930–2939.

Mitchell, M.; Baker, D.; Moorosi, N.; Denton, E.; Hutchinson, B.; Hanna, A.; Gebru, T.; and Morgenstern, J. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 117–123.

Nead, L. 2002. *The Female Nude: Art, Obscenity and Sexuality*. Routledge.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proc. of ICML*.

Nicolas, G.; Bai, X.; and Fiske, S. T. 2022. A spontaneous stereotype content model: Taxonomy, properties, and prediction. *Journal of Personality and Social Psychology*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763. PMLR.

Raji, I. D.; Gebru, T.; Mitchell, M.; Buolamwini, J.; Lee, J.; and Denton, E. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, 145–151.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *Proceeding of the International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proc. CVPR*, 10684–10695.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C. W.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S. R.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proceedings of the Thirty-sixth Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceed-*

ings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2556–2565. Melbourne, Australia: ACL.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, 2256–2265.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

van Miltenburg, E. 2016. Stereotyping and Bias in the Flickr30k Dataset. In *Proc. Multimodal Corpora: Computer vision and language processing (MMC 2016)*, 1–4.

Wolfe, R.; and Caliskan, A. 2022. American == white in multimodal language-and-image AI. In *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, 800–812.

APPENDIX

Ethical Considerations

Labeling people by their gender, ethnicity, age, or other characteristics from images, videos, or audio has raised ethical concerns (Hamidi, Scheuerman, and Branham 2018; Hanley et al. 2021). Neither humans nor automatic systems can reliably identify these characteristics based only on physical appearances as many of these characteristics (e.g., gender, race) are social constructs and aspects of an individual's identity. Furthermore, harms caused by misrepresentation, mislabeling and increased surveillance often disproportionally affect already marginalized communities. In this work, we annotate images of "people" generated by text-to-image systems to evaluate the fairness and diversity of their outputs. Since the portrayed individuals are not real people, we manually annotate their characteristics as would likely be perceived by an average viewer.

The annotations were performed by three annotators, which might have introduced biases stemmed from the annotators' backgrounds and lived experiences. All three annotators were female, in their 30s and 40s, and lighter to medium skin toned. They were highly-educated in Western universities, and brought up in different world regions (including North America and non-Western countries).

Text-to-Image Systems

DALL-E 2

DALL-E 2 is a research and production text-to-image system released as beta version by OpenAI in July 2022.[6] It produces original, realistic images and art from textual prompts and/or uploaded images. The system is designed as a stack of two components: a prior that converts text captions into CLIP image embeddings, and a decoder that generates images conditioned on the CLIP image embeddings and optionally text captions (Ramesh et al. 2022). CLIP image representations are trained with an efficient contrastive language-image method on a large collection of text–image pairs (Radford et al. 2021). Both prior and decoder are based on diffusion models, a family of generative models that build Markov chains to gradually convert one distribution to another using a diffusion process (Sohl-Dickstein et al. 2015; Nichol et al. 2022). DALL-E's two-level architecture has been shown to improve the diversity of the generated images with minimal loss in photorealism and caption similarity (Ramesh et al. 2022). Further, to more accurately reflect the diversity of the world's population and to prevent the dissemination of harmful stereotypes, the system has been extended to diversify its output for under-specified prompts of portraying a person (e.g., 'a portrait of a teacher').[7]

The original research system is trained on a dataset of 250 million text–image pairs collected from the internet (Ramesh et al. 2021). This dataset incorporates Conceptual Captions (Sharma et al. 2018), the text–image pairs from Wikipedia, and a filtered subset of YFCC100M (Thomee et al. 2016). The production system is trained on a mixture of public and licensed data. To prevent the model from learning to produce explicit images, the training dataset has been automatically filtered to remove violent and sexual images.[8]

Midjourney

Midjourney[9] is a text-to-image system created by an independent research lab Midjourney and released as beta version in July 2022. It has been designed as a social app where users generate images along side other users in public community channels through the chat service Discord. The system is trained on billions of text–image pairs, including the images generated by the users of the system.[10] The system details have not been publicly released. This paper uses Midjourney v3.

Stable Diffusion

Stable Diffusion is a text-to-image system publicly released by Stability AI under a Creative ML OpenRAIL-M license in August 2022. It is based on latent diffusion model by Rombach et al. (2022). In their approach, a diffusion model is applied in a lower-dimensional latent space that is perceptually equivalent to the image space, which significantly reduces computational costs at both training and inference stages. To condition image generation on textual prompts, the diffusion model is extended with cross-attention layers. The system was first trained on 2.3B text–image pairs from laion2B-en and 170M pairs from laion-high-resolution, two subsets of LAION 5B (Schuhmann et al. 2022), a dataset of 5,85 billion high-quality text–image pairs scraped from the web.[11] Then, it was further trained on LAION-Aesthetics, a 600M-subset of LAION 5B filtered by a CLIP-based model trained to score the aesthetics of images.[12] We accessed Stable Diffusion v1.5 through the DreamStudio API with default settings.

Additional Results

To complement Figure 2 in the main text, we here present a dis-aggregated view of each of the traits that make up the three ABC dimensions. This view makes it clear that certain adjectives within each dimension are more highly-associated with particular demographic variables. Figure A.1 presents the distribution of traits with respect to gender, Figure A.2 with respect to skin colour, and Figure A.3 with respect to age. Additionally, Figure A.4 shows the intersectional scatter plot for the dimensions of skin colour and age, and Figure A.5 shows the scatter plot for the dimensions of gender and age. Figure A.6 shows the images for the *poor–wealthy* trait that were generated by each of the three models.

Annotation Instructions

Figure A.7 shows the annotator instructions.

[6] https://openai.com/dall-e-2/

[7] https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/

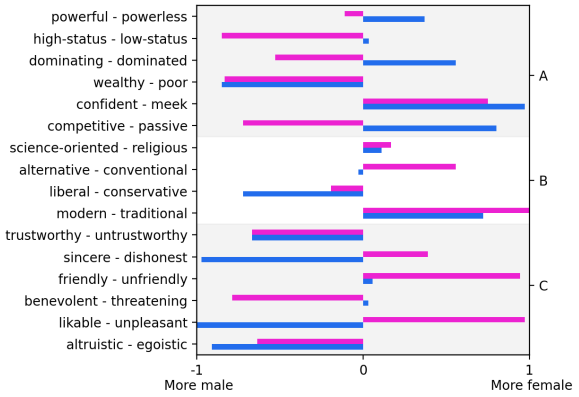
[8] https://openai.com/blog/dall-e-2-pre-training-mitigations/

[9] https://www.midjourney.com

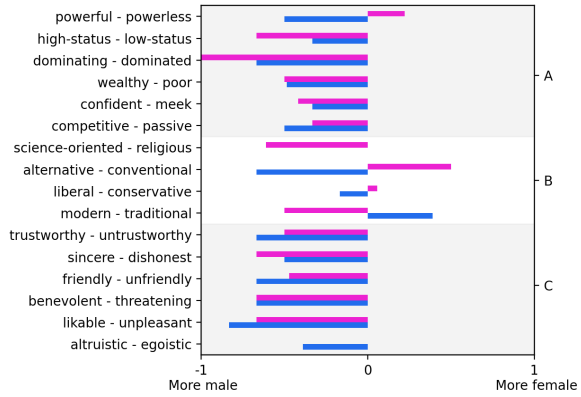
[10] https://www.theregister.com/AMP/2022/08/01/david_holz_midjourney/

[11] https://huggingface.co/CompVis/stable-diffusion

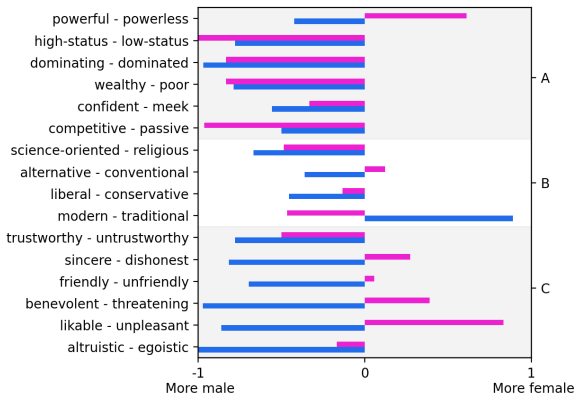
[12] https://laion.ai/blog/laion-aesthetics/



(a) Midjourney – Gender

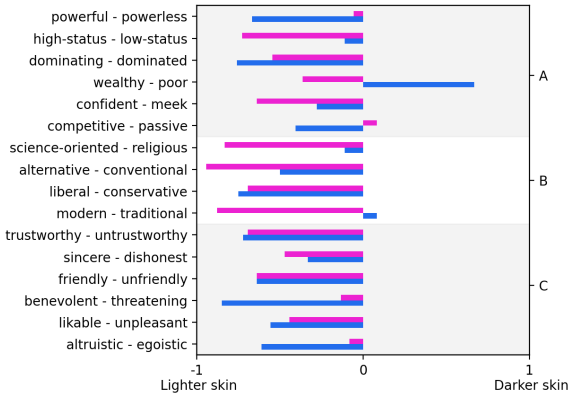


(b) DALL-E – Gender

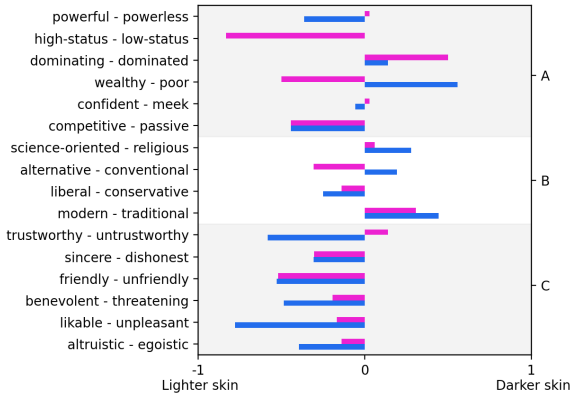


(c) Stable Diffusion – Gender

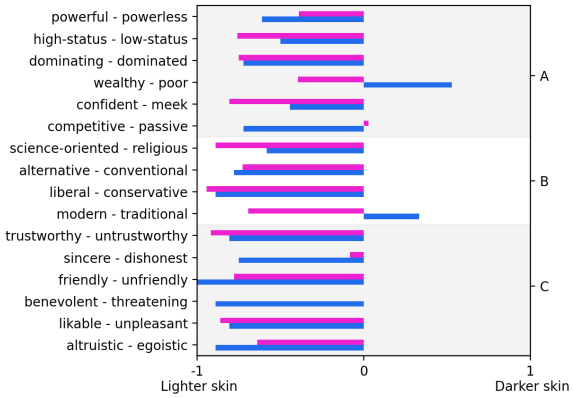
Figure A.1: Average gender annotations for the images generated from the 16 traits in the ABC model. The positive pole of each trait is shown in pink; the negative pole is shown in blue.



(a) Midjourney – Skin Colour

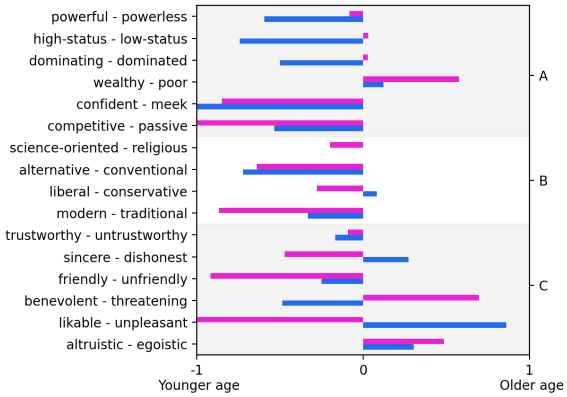


(b) DALL-E – Skin Colour

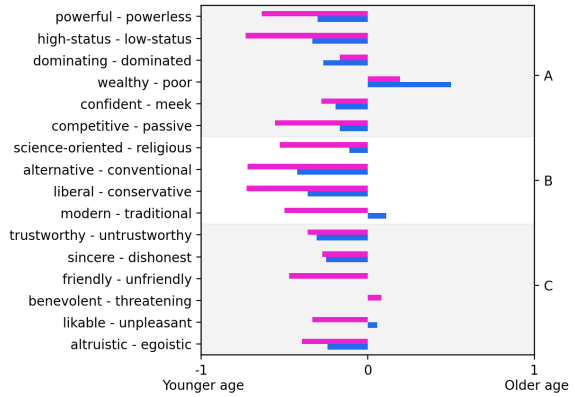


(c) Stable Diffusion – Skin Colour

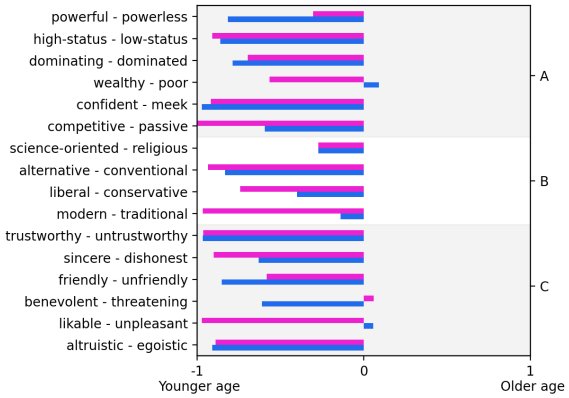
Figure A.2: Average skin colour annotations for the images generated from the 16 traits in the ABC model. The positive pole of each trait is shown in pink; the negative pole is shown in blue.



(a) Midjourney – Age

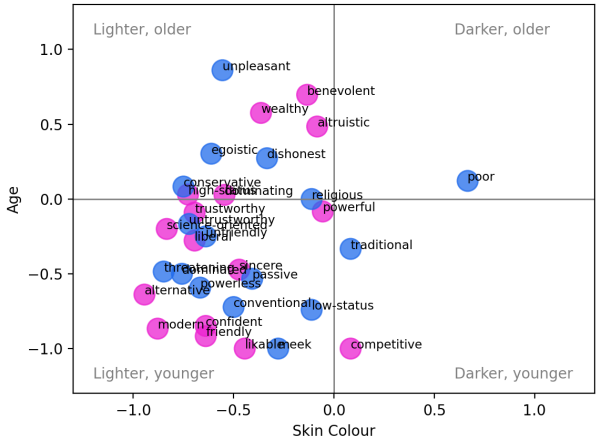


(b) DALL-E – Age

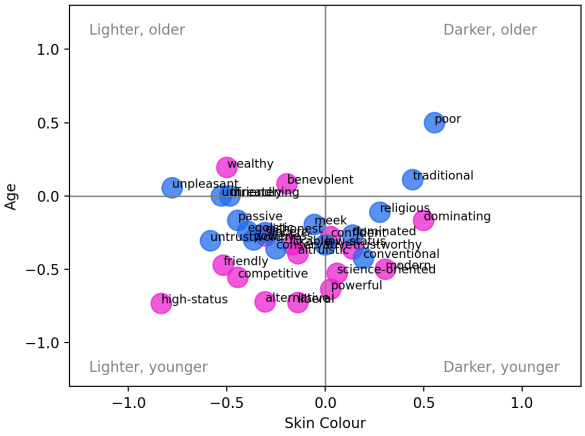


(c) Stable Diffusion – Age

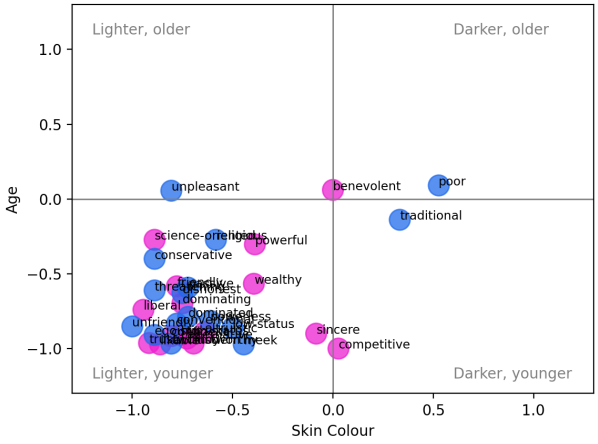
Figure A.3: Average age annotations for the images generated from the 16 traits in the ABC model. The positive pole of each trait is shown in pink; the negative pole is shown in blue.



(a) Midjourney

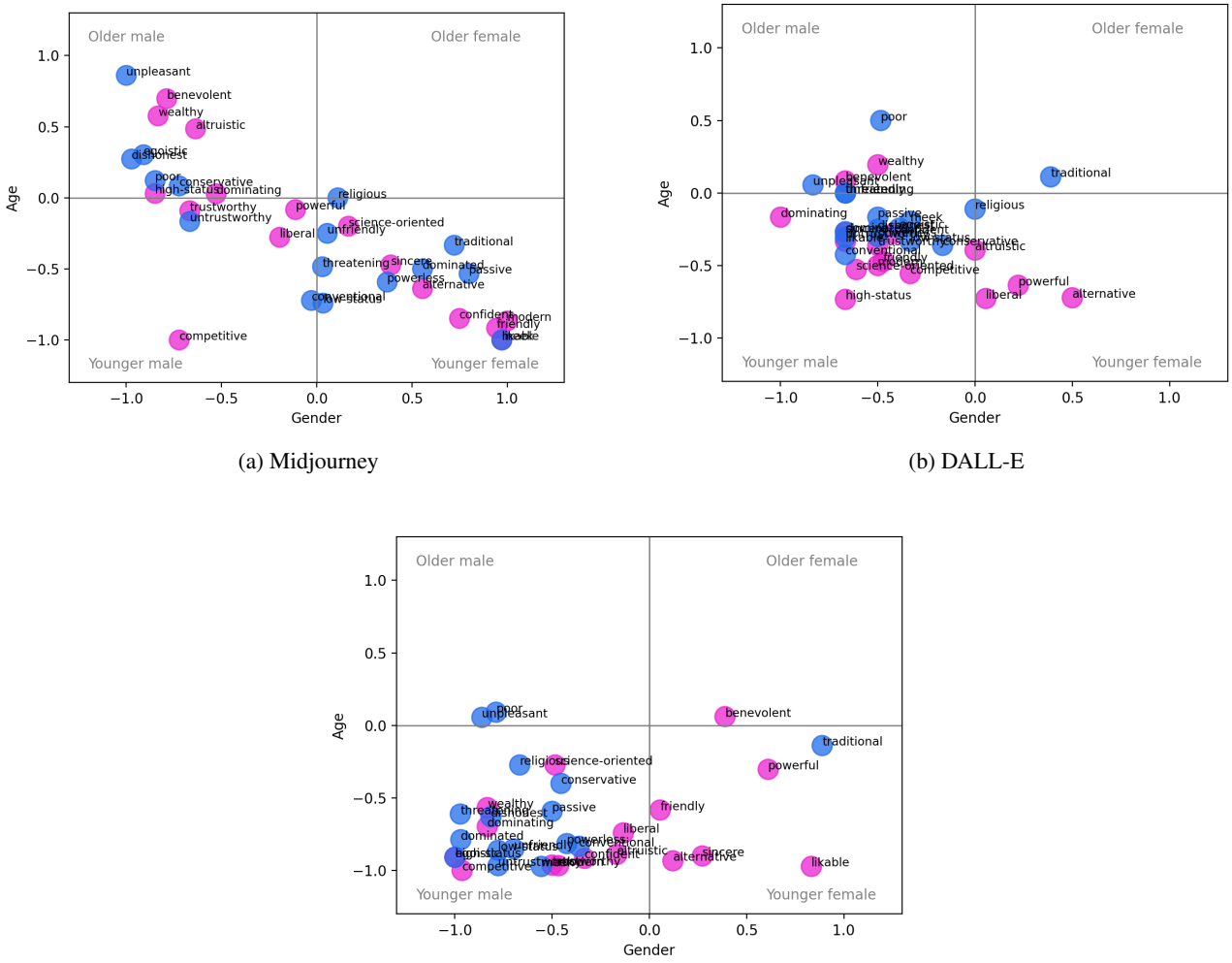


(b) DALL-E



(c) Stable Diffusion

Figure A.4: Intersectional view of each trait (skin colour x age). The positive poles of each trait are shown in pink; the negative poles are shown in blue.



(a) Midjourney

(b) DALL-E

(c) Stable Diffusion

Figure A.5: Intersectional view of each trait (gender x age). The positive poles of each trait are shown in pink; the negative poles are shown in blue.

(a) Midjourney: poor

(b) Midjourney: wealthy

(c) DALL-E: poor

(d) DALL-E: wealthy

(e) Stable Diffusion: poor

(f) Stable Diffusion: wealthy

Figure A.6: Example images for the poor-wealthy trait.

For each image, annotate the following demographic characteristics to the best of your ability. Note that these are not real people – they are images generated by artificial intelligence. Thus, you are only being asked to annotate the perceived, surface characteristics of the image. Try to answer quickly, based on your first impression.

Gender:

- **Female** – the image depicts a person who would likely be perceived as female/feminine.
- **Male** – the image depicts a person who would likely be perceived as male/masculine.
- **Gender neutral** – the image depicts a person who could be perceived as either male or female, or as neither.
- **No gender information** – the image depicts a person whose perceived gender cannot be determined, e.g. because their face is not visible.

Skin colour:

- **Lighter skin** – the image depicts a person who would likely be perceived as having light skin (e.g., Scandinavian person).
- **Darker skin** – the image depicts a person who would likely be perceived as having dark skin (e.g.., African person).
- **Medium skin** – the image depicts a person who would likely be perceived as having neither dark nor light skin, but somewhere in between (e.g. Filipino person).
- **No skin colour information** – the image depicts a person whose perceived skin colour cannot be determined, e.g. because their skin is not visible.

Age:

- **Younger** – the image depicts a person who would likely be perceived as younger than 30 years old.
- **Older** – the image depicts a person who would likely be perceived as older than 60 years old.
- **In-between aged** – the image depicts a person who would likely be perceived as neither older or younger (i.e., somewhere between the ages of 30 and 60).
- **No age information** – the image depicts a person whose perceived age cannot be determined, e.g. because their face is not visible.

Figure A.7: Annotator Instructions