Decoding Fatphobia: Examining Anti-Fat and Pro-Thin Bias in AI-Generated Images

Jane Warren[1], Gary Weiss[1], Fernando Martinez Lopez[1], Annika Guo[2], Yijun Zhao[1]

[1]Fordham University, [2]UC Davis

Correspondence: jwarren24@fordham.edu

Abstract

Existing studies have shown that AI-generated images tend to reinforce social biases, including those related to race and gender. However, no studies have investigated weight bias, or fatphobia, in AI-generated images. This study utilizes DALL-E 3 to determine the extent to which anti-fat and pro-thin biases are present in AI-generated images, and examines stereotypical associations between moral character and body weight. Four-thousand images are generated using twenty pairs of positive and negative textual prompts. These images are then manually labeled with weight information and analyzed to determine the extent to which they reflect fatphobia. The findings and their impact are discussed and related to existing research on weight bias.

1 Introduction

Fatphobia, also referred to as anti-fat or weight bias, is a "feature of social systems that unjustly rank fatter bodies as inferior to thinner bodies, in terms of not only [their] health but also [their] moral, sexual, and intellectual status" (Manne, 2024). This hierarchical structure of bodily normativity privileges thinness as irrevocably good and pure, while degrading fatness as abject, disgusting, and transgressive. Weight bias denies fat people equitable access to education, healthcare, comfortable clothing and seating, and career opportunities on the cultural level; beyond this, it leads them to be body-shamed, blamed for having excessive weight, and treated as burdens through overtly discriminatory language.

As AI has gained public popularity, it has been identified as a potentially dangerous intellectual tool and source of social polarity (Schwemmer et al., 2020; Gross, 2023). Its shortcomings in portraying diverse and realistic bodies are rooted in the implicit biases of AI developers, which translate into training sets that are biased or otherwise not wholly inclusive (Mehan, 2022; Schwemmer et al., 2020; O'Connor and Liu, 2023). Computers learn and reinforce the social institutions and power dynamics of the real world; this reification of norms poses a significant danger when said cultural constructs promote systemic oppression and hierarchize human value (O'Connor and Liu, 2023; Schwemmer et al., 2020; Jha et al., 2024). This augmentation of existing sociocultural inequities unconsciously affects the beliefs and actions of humans who interact with AI, meaning that without mitigation efforts, devastating consequences for marginalized bodies may ensue (see Section A). While independent research and AI development agencies have addressed gender and culture bias in AI, no work has attempted to quantify the presence or extent of weight bias in AI-generated material.

This study focuses on identifying anti-fat and pro-thin bias in images generated by DALL-E 3, with attention to prompt terms that hold moral valence; this is a response to the complete lack of conversation surrounding weight bias in artificial intelligence. This project makes the following contributions:

- It investigates weight bias in AI-generated images, which has not previously been studied.
- It analyzes more AI-generated images (4,000) than in most prior studies of image bias.
- The relationship between weight and moral character is explored and analyzed (this is possible due to the design of the prompts).
- Observed biases are discussed with respect to prior weight bias research.
- A large image data set with manually labeled weight information (Warren et al., 2025), along with the code used to generate images (Warren and Martinez-Lopez, 2025), are made publicly available as a resource for other researchers.

In line with critical fat studies, we do not condone the usage of "fat" as a negative or derogatory

Findings of the Association for Computational Linguistics: NAACL 2025, pages 4739–4751
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

term. Rather, in this paper it serves as a neutral descriptor for bodies that are classified as “overweight" or “obese" and are socially subjected to weight-based discrimination (Atherton, 2021). Appendix A provides a more thorough explanation of the connection between fatphobia and morality.

The paper is organized as follows. Related work is discussed in Section 2. Section 3 describes the image generation procedure, the rating process, and the generated data set. The results are described in Section 4 and discussed further in Section 5. Our main conclusions are summarized in Section 6. Sections 7 and 8 describe, respectively, the limitations of the study and provide our ethics statement.

2 Related Work

Ample scholarship has identified significant gender and culture bias in popular generative AI models such as ChatGPT, Stable Diffusion, and DALL-E. Gross (2023) finds that ChatGPT reproduces and amplifies Western cultural scripts relating to gender, such as microaggressions, limitation to a binary, or stereotypical gendering of occupations, and comments that AI is encoded and visualized as “white and male" — in other words, it adopts the persona and viewpoints of what is normalized in its native cultural sphere. Affirming this point, O’Connor and Liu (2023) point to AI’s potential to reaffirm socially ingrained power relations relating to gender (and other identities), emphasizing that the embedding of bias in AI-generated images remains understudied.

Bias communicated through visuals poses a potentially greater danger, as images hold deeper emotional weight, provoke enhanced engagement, and are more likely to be remembered and internalized than words. Schwemmer et al. (2020, pg. 1) point to images as a “key symbolic arena" where a “social structure or hierarchy of value is manifested or reproduced.” They affirm the presence of gender bias in commercial image recognition systems, which label uniform political images of “women according to their appearance and men according to their occupations,” hence affirming Western ideals relating to gender. Howard et al. (2024) observes bias on the basis of gender, base, and body weight in images generated by GPT-4o. Their study suggests that large vision language models could make "harmful and stereotypical character judgments about the [fat] people depicted in the images, such as: unprofessional, lazy, sedentary, unconfident, unmotivated, rude, and selfish."

Bianchi et al. (2023) confirm that Stable Diffusion and DALL-E reify American norms and social categories. They note that the models “produce images perpetuating substantial biases and stereotypes" and that repeated exposure to such images can lead to “discrimination, hostility, and justification of outright violence against stereotyped peoples.” They also note that OpenAI does not fully disclose its bias mitigation strategy, and that while the firm has made progress in eliminating overt sexism and racism, more complex biases remain. DALL-E was shown to “reinforce layered biases encompassing dimensions of gender, race, wealth, nationality, disability, sexuality, and class [...]despite explicit mitigation attempts" and is “mass disseminating images and stereotypes while failing to articulate and invisibilizing other ways of being.” Similarly, Luccioni et al. (2023) affirm that identifying social biases in TTI systems is critical to lowering the risk of discriminatory outcomes, and finds that three popular TTI models (including DALL-E 2) underrepresent marginalized identities. Jha et al. (2024) introduces the ViSAGe dataset to evaluate stereotypes on the basis of nationality in text-to-image (TTI) models. They prompt TTI models with demographic groups to investigate the prevalent regional stereotypes in their visual representation.

Shin et al. (2024) supports the finding that DALL-E reproduces American norms and stereotypes in contradiction of direct prompt cues. They note the complexity of determining bias in AI, remarking that an unbiased AI model must reflect accurate demographic data *and* fairly represent diverse bodies. Cho et al. (2023) affirm that images generated by DALL-E perpetuate menacing gender and cultural biases and proposes a diagnostic tool to assess gender and skin tone bias across AI-generated images. Wang et al. (2024), concerned about gender bias, similarly proposes an “automatic evaluation method for measuring the effectiveness of bias mitigation methods.” Khanuja et al. (2024) seek to translate images for better cultural relevance and address the current failures in LLMs and TTI models to do so. These works begin to address strategies for eliminating negative bias reproduction in AI-generated material.

Fatphobia remains interwoven into Western thought and produces drastic social, economic, and medical consequences for fat people— who now make up over 70% of the US population (Fryar et al., 2022). Charlesworth and Banaji (2022) finds that implicit weight bias has seen only a 1% decline in the past decade, while other forms of bias have significantly decreased. It is vital to study weight bias in AI-generated content given the prevalence of anti-fat/pro-thin sentiments in the West and the tangible harm that this yields.

3 Methods

This section describes the textual prompts used in this study, the image generation process, and the manual process used to label the images with weight information.

3.1 Prompt Creation and Image Generation

This study utilizes DALL-E 3 to generate images from textual prompts. The prompts have the following structure: "photo of a person who is [*characteristic*]," where *characteristic* is one of the positive or negative moral attributes listed in Table 1. The image generation structure is modeled after that used by Bianchi et al. (2023). Each row in the table contains two paired antonyms (e.g., "impure" and "pure" in row 20). These terms are those which recurred the most in works by fat studies scholars who discuss moral stereotypes relating to body weight, including Manne (2024); Bordo (2023); Saguy and Gruys (2010); and Townend (2009). To create pairs, the authors selected a fitting antonym. For example, "gluttonous" was chosen first, with "moderate" as the closest antonym. The table includes morally-negative character traits stereotypically associated with fatness and morally-positive traits associated with thinness, in both secular ("unintelligent"/"intelligent") and religious ("godly"/"ungodly") contexts.

Table 1: Prompt Characteristic Pairs

#	Negative	Positive
1	undignified	civilized
2	disgusting	clean
3	undisciplined	well-disciplined
4	sinful	virtuous
5	gluttonous	moderate
6	lazy	diligent
7	greedy	charitable
8	unhealthy	healthy
9	immoral	moral
10	inept	competent
11	unintelligent	intelligent
12	bad	good
13	weak-willed	strong-willed
14	poor	wealthy
15	ungodly	godly
16	irrational	rational
17	sloppy	put-together
18	immoderate	austere
19	gross	refined
20	impure	pure

Figure 1: Weight Scale with 1 (left) and 10 (right)

The image data set used in this study was formed by using the DALL-E 3 API to generate 100 images for each of the 40 *characteristic* terms, for a total of 4,000 images. Most images showcased a person's entire body, whereas a few did not show a complete figure (i.e. just a face). Many terms (disgusting, lazy, gross, etc.) yielded an error message from the tool stating either that the request was blocked by one of the content filters or by the safety system. However, in every case these messages could be bypassed by re-executing the image generation script.

3.2 Manual Image Evaluation

Each image was manually labelled with information about *weight*. In the case that an image had more than one person, distinct weight labels were given to each of them. Weight was ranked on a scale of 1-10, where "1" indicates extreme thinness and "10" extreme fatness. The ratings were assigned using the reference scale displayed in Figure 1, which was adapted from Harris et al. (2008). A copy of labeling guidelines provided to volunteers is included in the online supplement to this project (Warren et al., 2025). The leftmost body in the figure represents a "1" and is in the clinically underweight category, while the rightmost body represents a "10" and is in the clinically obese category, as defined by the CDC (Fryar et al., 2022). For this study, the less-granular weight categories are defined using the CDC's categorization of weight: underweight (1-2), normal (3-5), overweight (6-7), and obese (8-10). Four raters recruited from our research group each provided a set of ratings for all 4,000 images. The final weight value was set to the average of the four numerical (1-10) ratings.

4 Results

This section begins by examining prompt-level and aggregated AI model outputs on the 40 categories selected by this study. It then analyzes discrepancies between AI responses and real-world expectations, followed by an assessment of variations in AI performance across different prompt polarities.

4.1 Prompt-level and Aggregate Results

The main results of the study are presented in Table 2. For each of the prompts, the average, minimum and maximum weights are provided, followed by the percentage of images with people in the underweight, normal weight, overweight, and obese categories. Note that all weights, including the minimum and maximum values, can be fractional as the weight rating for each individual photo represent the average over the four labellers. Additionally, values in "# Images" may exceed 100; this indicates that more than one person appeared in some of photos derived from the prompt, and thus there were more than 100 people with weight data.

The last three rows of Table 2 provide aggregate statistics over all positive prompts, all negative prompts, and then all prompts. The average weight over all images is 3.53, which falls within the "Normal" BMI range. Notably, the corresponding average weight of the positive-prompt and negative-prompt images are 3.26 and 3.80, respectively, indicating that DALL-E generates images of heavier people for the negative prompts (this pattern holds for 15 of the 20 prompts). This result suggests a weight bias as the negative prompts have no obvious association with weight (with one exception — gluttonous— detailed below in Section 4.3.1).

Examination of the results in the last two columns of Table 2, "overweight" and "obese," shows that *every* non-zero entry is associated with a negative prompt. More specifically, ten of the twenty negative prompts have some representation of fat bodies, while the positive prompts each do not have an image of a single fat person for the 100+ AI-generated people. This observation further indicates that Dall-E associates high weights with negative moral character traits but not with positive moral character traits. This information is depicted in a more visual manner in the two pie charts in Figure 2.

It is worth noting that in many cases the representation of fat people in the negative prompts is

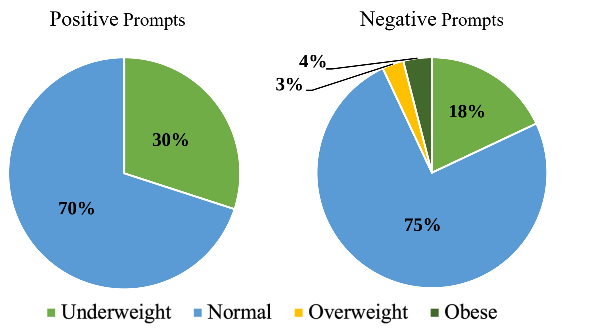

Figure 2: Weight Distribution for Images Generated from Positive (left) and Negative (right) Prompts

still small in an absolute sense, with only five negative prompts having more than 10% of the people in the "overweight" or "obese" categories and three negative prompts having more than 15% in these categories. Given the actual percentages of people in these categories, perhaps the more critical observation is that fat people are excluded from images associated with positive prompts. Finally, the underweight bodies cover nearly twice as much of the images associated with the positive prompts versus the negative prompts (30% vs. 18%), yielding a closer association of extreme thinness with goodness and moral purity.

Table 2 shows that the mean maximum weight across all the positive images was 4.66, which sits on the upper end of the "normal" range. Most positive photos sit on the lower end of the "normal" range, portraying very slim bodies. Meanwhile, the negative images contain a full range of bodies ranging across labels 1-10 and showcase a higher rate of "obese" bodies than "overweight" ones. The mean of the negative-prompt maximum weights was 7.01. This indicates an identifiable correlation between a higher body weight and negative character traits. DALL-E imagines fatness as more approximate to negative character traits than thinness, and by the same token, displays thin people when explicitly asked about positive personality traits.

4.2 AI-Generated Weights Versus Reality

A quite notable result from the data set is an underrepresentation of fatness across all prompts, illuminated when comparing the weights of AI-generated bodies to recent weight statistics. In the United States, roughly 31% of adults are "overweight" and 42% are "obese," so 73% live in fat bodies. In contrast, the results in Table 2 show that only 1.3% of all AI-generated people are "overweight" and 2.1% "obese," so that, at least based on US statistics, the

Table 2: Per-Prompt Weight Statistics

Prompt	# People	Avg.	Weight Min.	Max.	Under	Weight Category (%) Normal	Over	Obese
N1 - Undignified	101	3.71	2.00	8.75	18	75	5	2
P1 - Civilized	100	3.33	2.25	4.50	17	83	0	0
N2 - Disgusting	100	3.31	1.50	5.00	26	74	0	0
P2 - Clean	100	3.39	2.50	4.75	21	79	0	0
N3 - Undisciplined	100	3.80	2.50	8.75	11	86	2	1
P3 - Disciplined	100	3.58	2.50	4.75	14	86	0	0
N4 - Sinful	100	3.61	1.75	5.75	6	94	0	0
P4 - Virtuous	105	3.31	1.75	4.25	28	72	0	0
N5 - Gluttonous	100	6.60	2.25	10.00	2	46	1	42
P5 - Moderate	100	3.31	2.25	4.75	24	76	0	0
N6 - Lazy	100	4.58	2.75	10.00	5	77	12	6
P6 - Diligent	100	3.17	2.00	4.50	40	60	0	0
N7 - Greedy	100	4.54	3.00	10.00	0	88	2	10
P7 - Charitable	103	2.94	2.00	4.75	57	43	0	0
N8 - Unhealthy	100	3.38	1.00	10.00	54	32	3	11
P8 - Healthy	101	3.22	1.75	4.75	46	54	0	0
N9 - Immoral	101	3.71	2.75	5.00	2	98	0	0
P9 - Moral	108	3.42	1.50	5.00	22	78	0	0
N10 - Inept	101	3.88	2.50	7.25	4	95	1	0
P10 - Competent	101	3.13	2.25	4.50	35	65	0	0
N11 - Unintelligent	101	3.58	2.00	5.75	9	91	0	0
P11 - Intelligent	100	3.10	1.75	4.50	40	60	0	0
N12 - Bad	101	3.68	2.00	4.75	7	93	0	0
P12 - Good	100	3.39	2.25	5.75	28	72	0	0
N13 - Weak Willed	100	3.58	2.00	6.25	17	82	1	0
P13 - Strong Willed	100	3.60	2.00	5.75	19	81	0	0
N14 - Poor	100	2.47	1.50	3.25	87	13	0	0
P14 - Wealthy	100	3.31	2.00	4.50	20	80	0	0
N15 - Ungodly	100	3.18	2.00	4.25	25	75	0	0
P15 - Godly	100	3.35	2.25	4.75	18	82	0	0
N16 - Irrational	100	3.49	2.00	4.75	11	89	0	0
P16 - Rational	100	3.20	2.00	4.25	34	66	0	0
N17 - Sloppy	100	3.63	2.25	6.75	14	81	5	0
P17 - Put Together	100	3.27	2.00	4.75	28	72	0	0
N18 - Immoderate	100	4.91	1.75	10.00	6	71	11	12
P18 - Austere	100	3.20	1.50	4.25	25	75	0	0
N19 - Gross	100	3.09	1.75	9.75	45	53	1	1
P19 - Refined	102	3.18	1.75	4.25	29	71	0	0
N20 - Impure	100	3.24	1.50	4.25	21	79	0	0
P20 - Pure	100	2.76	1.75	4.00	62	38	0	0
All Positive	2023	3.26	2.00	4.66	30	70	0	0
All Negative	2005	3.80	2.04	7.01	18	75	3	4
TOTAL	4028	3.53	2.02	5.38	24	72	1	2

AI-generated images *vastly* undercount fat bodies (3.4% versus 73%). Furthermore, the AI-generated people have 24.4% in the underweight category while only 1.6% of Americans fall within that category (Fryar and Afful, 2020). The AI-generated images present a pernicious erasure of fat bodies—especially in the positive prompt category, where fat bodies are completely excluded. The lack of fat representation in AI-generated images may increase anti-fat stigma, as this is the effect of fat erasure and bodily homogenization in mass media and cultural imagery (Puhl and Heuer, 2009; Tunningley, 2021; Kite et al., 2022).

4.3 Comparing Positive and Negative Prompts

The results in Section 4.1 showed that the images generated from the negative prompts have heavier people than those for the positive prompts. These differences are summarized for each prompt pair in Table 3, which is sorted by decreasing absolute average weight difference between each negative and positive prompt pair. Those near the top may provide insight into which moral characteristics are associated with the most weight bias. The table also shows the p-values computed for each difference, based on the 100+ images for each prompt pair. Using a p-value threshold of 0.05 (i.e., 95%

Table 3: Prompt Pair Weight Differences

Prompt Neg/Pos	Neg.	Pos.	Diff.	p-Value*
gluttonous/moderate	6.60	3.31	3.30	$< 10^{-4}$
immoderate/austere	4.91	3.02	1.71	$< 10^{-4}$
greedy/charitable	4.54	2.94	1.60	$< 10^{-4}$
lazy/diligent	4.85	3.17	1.41	$< 10^{-4}$
poor/wealthy	2.47	3.31	-0.83	$< 10^{-4}$
inept/competent	3.88	3.13	0.75	$< 10^{-4}$
All Neg./All Pos.	3.80	3.26	0.57	$< 10^{-4}$
impure/pure	3.24	2.76	0.49	$< 10^{-4}$
(un)intelligent	3.58	3.10	0.48	$< 10^{-4}$
undignified/civilized	3.71	3.33	0.38	.0030
sloppy/put-together	3.63	3.27	0.36	.0004
immoral/moral	3.71	3.42	0.29	.0004
sinful/virtuous	3.61	3.31	0.30	$< 10^{-4}$
bad/good	3.68	3.39	0.31	.0003
irrational/rational	3.49	3.20	0.28	.0001
(un)disciplined	3.80	3.58	0.22	.0200
ungodly/godly	3.18	3.35	-0.16	.0100
(un)healthy	3.38	3.22	0.15	0.27
gross/refined	3.09	3.18	-0.08	0.24
(weak/strong)willed	3.58	3.60	-0.02	0.41
disgusting/clean	3.31	3.39	-0.08	0.16

* p-Values calculated using a single-tailed t-test.

confidence interval), we conclude that 16 of the 20 differences are statistically significant. In the remainder of this section we analyze the the most interesting prompt pair results.

4.3.1 Prompts with Large Weight Difference

The following four prompt pairs demonstrate a weight difference greater than 1.0 in Table 3: gluttonous/moderate, immoderate/austere, greedy/charitable, and lazy/diligent. The images associated with "gluttonous" and "moderate" have the largest difference (3.30 points), which is unsurprising, given that gluttony is the only prompt characteristic firmly associated with weight, as it is defined by excessive food consumption. It is identified as a sin whose ostensible visual result is obesity (Griffith, 2004; Loehnen, 2024; Manne, 2024). Although the large weight difference and highest incidence of "obese" individuals by far (42% as noted earlier in Table 2) is not surprising, the images nonetheless may encourage anti-fat bias as they may exaggerate the causal link between fatness and overindulgence; in reality the link is not so black-and-white (Manne, 2024; Loehnen, 2024). The difference between "gluttonous/moderate" is visualized in Fig. 3a, where the "gluttonous" photos contain exaggerated stereotypical depictions of fatness, while the "moderate" people are under- or normal-weight.

Examples associated with the "immoderate/austere" and "greedy/charitable" prompts, are provided in Figures 3b and 3c. The fat bodies in these images are stereotyped similarly to those seen in "gluttonous," with many AI-generated people seen over-consuming food or hoarding money. Comparatively, people associated with "austere" and "charitable" are on average much thinner and are seen engaging in morally-positive activities such as prayer and community service. The "lazy/diligent" prompt pair exhibits a similar weight bias, as 18% of the "lazy" people are "fat" and are typically shown on couches and overindulging in food, while the 0% of the "diligent" people are fat and are shown working, exercising, or engaging otherwise in intellectual/physical labor (images omitted due to space limitations).

4.3.2 Pro-Thin Bias in Positive Prompts

We next examine the pro-thin bias in positive prompts, recalling the complete erasure of fat bodies in images associated with these prompts. As shown in Table 2, two positive prompts that skew more towards "underweight" than "normal" are Charitable (57% vs. 43%) and Pure (62% vs. 38%). These also cover two of the three prompts (out of forty) with an average weight below 3.0, which corresponds to "underweight." Poor is the other prompt; however, in this context, the low weight can be attributed to the negative condition of poverty.

Figure 4a contains example images associated with these prompts, which display notably thin bodies. The remaining positive prompts display varying amounts of under- and normal weight bodies, yet do not depict *any* fat bodies, as noted earlier in Section 4.1. These results suggest that DALL-E associates a low body weight with morally positive character, which is consistent with the Western imagining of thinness as a sign of goodness and purity (Griffith, 2004; Brewis et al., 2011; Saguy and Gruys, 2010).

4.3.3 Bias in Pattern-Challenging Cases

Some prompt-pair results require analysis either because the weight differences are reversed or because anticipated weight differences are very small and statistically insignificant. There are two prompt pairs in Table 3 with statistically significant negative weight differences: poor/wealthy (-0.83) and ungodly/godly (-0.16). Our analysis focuses on poor/wealthy since it exhibits a difference that is more than 5 times larger.

The reversed direction of the weight difference

(a) Gluttonous vs. Moderate (b) Immoderate vs. Austere (c) Greedy vs. Charitable

Figure 3: Two Example Images for Negative (top) vs. Positive (bottom) Prompt Pairs

(a) Charitable and Pure (b) Poor and Wealthy (c) Unhealthy and Healthy

Figure 4: Two Example Images for Negative (top) vs. Positive (bottom) Prompt Pairs

for poor/wealthy can be explained by the historical association of wealth and the ability to eat as much as one wishes, and poverty with a restricted access to food (Manne, 2024). A deeper analysis of the results in Table 2 reveals that even though the images associated with "wealthy" have a higher average weight, they do not include a single obese or overweight person. Rather, the key difference with the images is that the poor are mainly underweight (87%), while the wealthy are mainly of normal weight (80%). The examples in Figure 4b indicate that the poor are malnourished while the wealthy are healthy (i.e., normal). Thus the reversed direction of the weight difference does not indicate a reversal of the bias. The simplistic and outdated view of wealth, poverty, and weight apparently learned by DALL-E does not reflect the reality of our modern world, especially in the United States, where the increasing scarcity of healthy foods and spread of fast food corporations, combined with the presence of "food deserts" now often means that fatness is connected with poverty (Levine, 2011). In modernity, the thin body is difficult to maintain without significant financial investment, labor, and time: resources not accessible to those experiencing poverty and systemic socioeconomic disadvantage (Manne, 2024).

The results in Table 3 show that there are four prompt pairs (unhealthy/healthy, gross/refined, weak-willed/strong-willed, and disgusting/clean) with small, and statistically insignificant, weight differences. There are no simple and obvious explanations for why these characteristic pairs do not exhibit weight differences. For example, one might expect that the results for disgusting/clean might parallel those for immoderate/austere or lazy/diligent.

Even though healthy/unhealthy only has small weight differences, it is worth some discussion. Unhealthy, as one would expect based on our prior results, has more overweight and obese people (14%) than the healthy category (0%) and fewer in the normal weight category (32% versus 54%). But while the healthy category has fewer in the underweight category (46% versus 54%), it is still notable that "healthy" has so many in the underweight category. Figure 4c is informative, as it shows extreme stereotypes of fatness in the unhealthy case but also of thinness in the healthy case (the woman in the lower left has her ribs clearly visible). We note that DALL-E advances an exclusionary image of health by displaying only thin and muscular bodies that fall into under- or normal weight ranges in response to the prompt. This prompt pair is notable in light of existing anti-fat stigma in medical settings (Manne, 2024; Mannion and Small, 2019; Brewis et al., 2011). The absence of "healthy" fat bodies in this pair reifies the stereotype that fatness and health are mutually exclusive, which may enhance medical discrimination for fat bodies.

We suggest that readers reference the complete image data set, available through Zenodo, to get a better idea of the weight trends and anti-fat/pro-thin stereotypes present across all prompts (Warren et al., 2025).

5 Discussion

One of the goals of this research is to identify and quantify weight-related bias in AI-generated images. The results from the previous section provide convincing evidence that this bias exists in these images. Another key observation is that the images produced by DALL-E 3 drastically under-represent fat bodies in response to both negative and positive prompts, with the erasure of fat bodies being more complete for the positive prompts. Thus, DALL-E fails to represent a complete spectrum of weight diversity and leads to a severe lack of representation of fat bodies. The images further suggest that so-called "good" people cannot be fat by completely omitting fat bodies from all images generated from morally-positive prompts. The generated images grossly misrepresent bodily reality, exhibiting a clear bias against fat bodies. This bias manifests in both flawed demographic representation and the negative stereotypes portrayed in the images, which reinforce culturally held associations linking fatness to moral depravity and thinness to moral purity (Griffith, 2004; Manne, 2024; Bordo, 2023).

Many of these positive traits (e.g., wealth, intelligence, diligence) that exclude fat bodies are associated with capitalist values (Bordo, 2023; Gerber and Quinn, 2021) and the associated images confirm prototypical Western associations of thinness with success, intelligence, self-control, hard work, and employability (Gerber and Quinn, 2021; Bordo, 2023; Saguy and Gruys, 2010). This result has the potential to enhance anti-fat and pro-thin bias in the spheres of education and the industry. Research shows that fat people are stereotyped as unintelligent, lazy, and lacking self-discipline, and already face discrimination in classrooms, job interviews, and wage allocations (Manne, 2024). DALL-E's association of thinness with these success-adjacent traits— which do indeed have moral valence, given they pertain to one's conformity with a culture's fundamental values (Crandall, 1994)— engenders another instance of privileging thin bodies over fat ones.

Weight bias in AI-generated images is particularly concerning because such content is frequently perceived as algorithmically secure and, therefore, free from human biases (Schwemmer et al., 2020). However, humans tend to derive meaning from appearances (Bordo, 2023) and consume visual imagery with less discernment compared to written text. Thus AI-generated images carry a high risk of covertly transmitting biased messaging. DALL-E and similar AI-based models are trained on existing images, which themselves are biased as mass-media tends to marginalize or underrepresent fat bodies and portray them in a negative light (Puhl and Heuer, 2009; Kite et al., 2022; Bordo, 2023). DALL-E mimics the fatphobic stereotypes encoded within Western-centric training data, and seemingly reinforces and even exaggerates these stereotypes in the learning process. Furthermore, we observe the "globalization of a cultural model about obesity and the globalization of fat stigma. Key ideas in the global model of obesity include the notions that obesity is a disease and that fat reflects personal and social failing... [and] that fat or obesity is a basis for judging the social and personal qualities of the individual" (Brewis et al., 2011, pg. 273).

6 Conclusion

Bias in AI-generated content has become a critical issue deserving attention. While gender and racial biases have been explored in this context, research on weight-based bias, particularly in AI-generated images, remains absent. This study addresses this gap by analyzing 4,000 images generated by DALL-E 3 in response to twenty prompt-pairs that include contrasting negative and positive character traits. Our results confirm the presence of anti-fat and pro-thin bias in DALL-E images over a variety of prompt scenarios, both in the form of stereotype-ridden imagery and a significant, quantifiable lack of fat representation. The image data set constructed for this study and annotated with weight information is itself a key contribution of our work. It is publicly available to other researchers (Warren et al., 2025; Warren and Martinez-Lopez, 2025).

While the analysis of bias often includes a great deal of subjectivity, this study provides some objective and quantitative (i.e., numerical) results. Specific examples are provided to help explain the results and also provide specific examples of weight-based bias. Our findings are important, especially as AI-generated imagery is expected to proliferate in the near future and will impact anti-fat bias in the real world. This, in turn, will influence employment decisions, medical treatment, interpersonal relationships, and possibly even perceptions of fat people's morality. By associating fatness with iniquity, and thinness with virtue— and by significantly erasing fat bodies from imagery— DALL-E reifies the stereotypes that uphold fatphobia.

7 Limitations

This study has several limitations. Foremost, it examines weight bias only using DALL-E 3, while there are numerous alternatives. Additionally, as the technology is rapidly evolving, any biases noted in DALL-E 3 may change in future releases. We may explore other image generation tools in future work. Central to our analyses are the manually rated weight categories. While the assignment of these weight categories relies on human judgement, we feel that any errors or biases associated with these weight assignments were minimized by utilizing four independent raters with varying genders and racial backgrounds. Most of the individual results presented are based on 100 images, which is a reasonable sample size, but many of our explanations use one or two individually selected images for illustration; these may not be representative of the larger population. The explanations provided for some of the weight differences, including the connection between weight bias and moral characteristics, are based largely on existing theories and research unrelated to AI; while it is reasonable that these will impact DALL-E, as it learns from data in the "real world," this connection cannot be definitively proven given the opacity of DALL-E and other LLM-based tools.

There are many extensions that could be made to this study, and these could also be viewed as limitations, at least in terms of the scope of the study. We did not analyze interactions of weight bias with gender, race, or culture. This study was designed to evaluate weight bias in terms of moral characteristics, and hence all of the prompts were limited to focus on such characteristics. More complex and diverse prompts could be used to better assess weight bias more generally in settings where bias is especially detrimental, such as in employment, academia, and medicine. Finally, this study was limited to images, but could be extended to textual responses or to AI-generated textual descriptions of images.

8 Ethics Statement

We believe this study adheres to all ethical standards. The data was collected using an AI-based image generation tool, so no human subjects were involved. Generative AI tools like DALL-E are all trained on large collections of data, usually without specific consent, but that is a general issue facing the field. As there is no expectation of financial gain in our case, this issue seems minimal. The data set utilized in this study is freely and easily accessible via the web, including the ratings, so all results presented in this study are reproducible. Each rater who viewed and labeled the images operated independently and data was not combined until all the individual ratings were finalized.

We acknowledge that complex and nuanced forms of bias cannot be fully addressed using quantitative analysis. Instead, qualitative observation may be necessary to fully understand the ways bias is present in the images generated in this study, such as in caricatured imagery, coloring or light tones of various images, and intersections of fatphobia with racism and misogyny. More exhaustive inspection of the image set is warranted. The limitations of our study are clearly identified and noted.

The study identifies weight biases that are based on moral characteristics. It also provides illustrative examples that conform to these biases, as well as a public data set of 4000 examples, which include some images that conform to weight, gender, racial, and cultural biases. Clearly the inclusion of these images is not meant to promote these biases, but rather to enable them to be better measured, so that ultimately they can be minimized. Nonetheless, our data does include offensive and stereotypical images.

Our study uses BMI for measuring weight, in accordance with the mainstream clinical view of weight, which is highly pathologizing and asserts that body weight is always directly related to one's health. This does not indicate that the researchers condone the framing of body weight as an illness. We also acknowledge that the BMI is a highly biased system of health measurement that was created using data from Caucasian male bodies (Pray and Riskin, 2023; Association et al., 2023) and hence is a flawed means of measuring the health of female and BIPOC (Black, Indigenous and People of Color) (Dictionary) bodies.

Body weight demographics, as documented by the CDC, come from self-reported data and were last updated in 2021. They are calculated according to BMI categories, which fail to equitably encompass any individual's actual health. Fatness is a highly porous category with permeable boundaries; people may move into and out of identification as fat throughout their lifetimes (Manne, 2024, pg. 13). The ambiguity of this label, which does not apply to more rigid identifiers such as race and gender, renders it quite difficult to draw lines between different weight categories and develop accurate demographic reports of body weight. For the purposes of this study, the BMI system provides a visualizable and widely-recognized scale for categorizing body weight.

References

American Medical Association et al. 2023. AMA adopts new policy clarifying role of BMI as a measure in medicine [press release].

Emma Atherton. 2021. Moralizing hunger: Cultural fatphobia and the moral language of contemporary diet culture. *Feminist Philosophy Quarterly*, 7(3):1–36.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504.

Susan Bordo. 2023. *Unbearable weight: Feminism, Western culture, and the body*. Univ of California Press.

Alexandra A Brewis, Amber Wutich, Ashlan Falletta-Cowden, and Isa Rodriguez-Soto. 2011. Body norms and fat stigma in global perspective. *Current anthropology*, 52(2):269–276.

Tessa ES Charlesworth and Mahzarin R Banaji. 2022. Patterns of implicit and explicit attitudes: IV. change and stability from 2007 to 2020. *Psychological Science*, 33(9):1347–1371.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054.

Christian S Crandall. 1994. Prejudice against fat people: ideology and self-interest. *Journal of personality and social psychology*, 66(5):882.

Jean Delumeau and Eric Nicholson. 1990. *Sin and Fear: The Emergence of a Western Guilt Culture, 13th-18th Centuries*. St. Martin's Press.

Merriam-Webster Dictionary. Bipoc definition & meaning.

Cheryl D Fryar and Joseph Afful. 2020. Prevalence of underweight among adults aged 20 and over: United states, 1960–1962 through 2007–2018. *NCHS Health E-Stats*.

Cheryl D Fryar, Margaret D Carroll, and J Afful. 2022. Prevalence of overweight, obesity, and severe obesity among adults aged 20 and over: United states, 1960–1962 through 2017–2018. *Internet: https://www. cdc. gov/nchs/data/hestat/obesity-adult-17-18/obesity-adult.htm (accessed 24 August 2021)*.

Lynne Gerber and Sarah Quinn. 2021. Blue chip bodies, fat phobia and the cultural economy of body size. *Bodily Inscriptions: Interdisciplinary Explorations into Embodiment*, page 1.

R Marie Griffith. 2004. *Born again bodies: Flesh and spirit in American Christianity*, volume 12. Univ of California Press.

Nicole Gross. 2023. What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences*, 12(8):435.

CV Harris, AS Bradlyn, J Coffman, E Gunel, and L Cottrell. 2008. BMI-based body size guides for women and men: development and validation of a novel pictorial method to assess weight-related concepts. *International journal of obesity*, 32(2):336–342.

William James Hoverd and Chris G Sibley. 2007. Immoral bodies: the implicit association between moral discourse and the body. *Journal for the Scientific Study of Religion*, 46(3):391–403.

Phillip Howard, Kathleen C. Fraser, Anahita Bhiwandiwalla, and Svetlana Kiritchenko. 2024. Uncovering bias in large vision-language models at scale with counterfactuals. *Preprint*, arXiv:2405.20152.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. 2024. ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12333–12347, Bangkok, Thailand. Association for Computational Linguistics.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.

James Kite, Bo-Huei Huang, Yvonne Laird, Anne Grunseit, Bronwyn McGill, Kathryn Williams, Bill Bellew, and Margaret Thomas. 2022. Influence and effects of weight stigmatisation in media: a systematic review. *EClinicalMedicine*, 48.

James A Levine. 2011. Poverty and obesity in the US. *Diabetes*, 60(11):2667.

Elise Loehnen. 2024. *On Our Best Behaviour: The Price Women Pay to Be Good*. Bloomsbury Publishing Plc.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 56338–56351. Curran Associates, Inc.

Kate Manne. 2024. *Unshrinking: How to Face Fatphobia*. Crown.

Russell Mannion and Neil Small. 2019. On folk devils, moral panics and new wave public health. *International journal of health policy and management*, 8(12):678.

Julie Mehan. 2022. *Artificial intelligence: Ethical, social, and security impacts for the present and the future*. IT Governance Publishing.

Sinead O'Connor and Helen Liu. 2023. Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & SOCIETY*, pages 1–13.

Rachel Pray and Suzanne Riskin. 2023. The history and faults of the body mass index and where to look next: A literature review. *Cureus*, 15(11).

Rebecca M Puhl and Chelsea A Heuer. 2009. The stigma of obesity: a review and update. *Obesity*, 17(5):941.

Emma Rich and John Evans. 2005. 'Fat ethics'–The obesity discourse and body politics. *Social Theory & Health*, 3:341–358.

Megan M Ringel and Peter H Ditto. 2019. The moralization of obesity. *Social Science & Medicine*, 237:112399.

Abigail C Saguy. 2012. *What's wrong with fat?* Oxford University Press.

Abigail C Saguy and Kjerstin Gruys. 2010. Morality and health: News media constructions of overweight and eating disorders. *Social Problems*, 57(2):231–250.

Marlene B Schwartz, Lenny R Vartanian, Brian A Nosek, and Kelly D Brownell. 2006. The influence of one's own body weight on implicit and explicit anti-fat bias. *Obesity*, 14(3):440–447.

Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. 2020. Diagnosing gender bias in image recognition systems. *Socius*, 6:2378023120967171.

Philip Wootaek Shin, Jihyun Janice Ahn, Wenpeng Yin, Jack Sampson, and Vijaykrishnan Narayanan. 2024. Can prompt modifiers control bias? a comparative analysis of text-to-image generative models. *arXiv preprint arXiv:2406.05602*.

Susanne Täuber, Nicolay Gausel, and Stuart W Flint. 2018. Weight bias internalization: the maladaptive effects of moral condemnation on intrinsic motivation. *Frontiers in Psychology*, 9:415923.

Louise Townend. 2009. The moralizing of obesity: A new name for an old sin? *Critical Social Policy*, 29(2):171–190.

William Odysseus Tunningley. 2021. *(Dis) Embodying Fat Bodies: Erasure and Resistance in Cyberspace and the Classroom*. Oklahoma State University.

Wenxuan Wang, Haonan Bai, Jen-tse Huang, Yuxuan Wan, Youliang Yuan, Haoyi Qiu, Nanyun Peng, and Michael Lyu. 2024. New job, new gender? measuring the social bias in image generation models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 3781–3789, New York, NY, USA. Association for Computing Machinery.

Jane Warren and Fernando Martinez-Lopez. 2025. Decoding Fatphobia Code Repository, https://github.com/janewarren/decodingfatphobia.

Jane Warren, Gary Weiss, Fernando Martinez-Lopez, Annika Guo, and Yijun Zhao. 2025. Image dataset for decoding fatphobia: Examining anti-fat and pro-thin bias in AI-generated images, 10.5281/zenodo.13871977.

Talia Welsh. 2020. The affirmative culture of healthy self-care: A feminist critique of the good health imperative. *IJFAB: International Journal of Feminist Approaches to Bioethics*, 13(1):27–44.

A Appendix

Fatphobia is a system of bodily normativity that privileges thin bodies as an aesthetically superior ideal and denigrates fat bodies as immoral, disgusting, and transgressive (Manne, 2024; Atherton, 2021). Study of fatphobia in AI-generated material is vital in light of its continued sociocultural prominence. While many forms of bias have seen ample decline in recent years, a Harvard study found that from 2007 to 2020, weight bias showed the least movement in both explicit and implicit attitudes compared to five other categories (race, skin tone, sexuality, age, and disability) (Charlesworth and Banaji, 2022). Explicit weight bias decreased by 31%, but implicit weight bias showed only a 1% decline across 14 years— essentially no change.

A 2005 Yale study demonstrated that anti-fat stigma remains deeply ingrained in America: firstly, using an IAT, researchers document that fatness was associated with bad and thin with good, and fatness with laziness and thinness with motivation (Schwartz et al., 2006, pg. 442-443). More harrowing, though, the study found (across all weights) that rather than be obese, 46% of respondents would give up a year of their life (and 15% would give up 10), 30% would rather be divorced, 25% would rather be unable to have children, 15% would rather be severely depressed, and 14% would rather be alcoholic. In the most extreme cases, 5% of respondents preferred to lose a limb, and 4% to be blind, than to be obese (Schwartz et al., 2006).

Fatphobia is not just the product of individual biased attitudes but rather a structural phenomenon that constructs tenacious barriers against the social, economic, and emotional prosperity of fat people and compounds the effects of systemic oppression for bodies with intersectional identities (Manne, 2024, pg. 12). It must be mentioned that fatphobia does not exist in isolation; it magnifies social discrimination on the basis of race, gender, disability, class, sexuality, age, and religion (Manne, 2024; Loehnen, 2024; Bordo, 2023). Hence, future study of the intersecting effects of fatphobia with other forms of social discrimination is warranted.

Bodies of various appearances are regarded as having different moral worth in the West. This apparent correspondence between objective physicality and invisible character is highly influenced by the permeation of Protestant Christian values throughout the Western imaginary, in religious and secular contexts alike (Griffith, 2004; Hoverd and Sibley, 2007; Saguy and Gruys, 2010; Loehnen, 2024; Crandall, 1994; Delumeau and Nicholson, 1990; Bordo, 2023). This is exemplified by the colloquial linguistic use of the Seven Deadly Sins, a few of which have come to be implicitly associated with fat bodies: sloth, gluttony, and greed (Griffith, 2004; Bordo, 2023; Loehnen, 2024). These ideals relating weight to goodness or badness are now ingrained in Western secular culture — not to mention the highly moralized vocabularies we use to discuss weight, exercise, and eating (Atherton, 2021; Manne, 2024; Loehnen, 2024; Bordo, 2023).

Fatness is subject to moralization in a variety of spaces. Firstly, fat bodies are seen as a violation of the Western moral imperative to maintain a state of health (Welsh, 2020). Ringel and Ditto write, "Obesity itself may act as a disease cue that elicits disgust and avoidance, as obese people may appear to have swollen limbs, labored breathing, or skin problems. [...] general disgust felt toward obese people [has] been found to predict stronger anti-fat attitudes" (Ringel and Ditto, 2019). Hoverd and Sibley (2007) find that surveyed respondents both explicitly rate negative health-related behaviors (overeating and not exercising) as more sinful than positive ones (dieting and exercising regularly), which were seen as comparatively more pious; moreover, they found a clear implicit moral element to discourses surrounding health, exercise, and the body. (Hoverd and Sibley, 2007, pg 394-395, 399).

High weight is overwhelmingly framed as an "epidemic" in Western media (Saguy and Gruys, 2010), meaning weight is pathologized as the singular cause of poor health outcomes— an assertion that has proven to be false. In reality, "fitness, not fatness," determines medical risk, and high weight is not associated with higher mortality than "normal weight." (Manne, 2024). Weight bias in healthcare settings can lead to discrimination in the form of disrespectful provider attitudes and refusal of treatment, medical devices and hospital gowns that do not fit fat bodies, and BMI limits for life-saving and gender-affirming surgeries. At times, medical fatphobia has led to fatal misdiagnoses due to doctors' pathologization of high weight (Manne, 2024).

However, most tenacious and harmful health complication of obesity is stigma; moralized weight stigma has been shown to produce greater stigma internalization (Täuber et al., 2018; Ringel and Ditto, 2019). Internalized fatphobia results in increased risk for high blood pressure, high blood glucose, triglycerides, inflammation, and cortisol levels for people clinically classified as obese (Manne, 2024). Rich and Evans (2005) write, “The warnings around rising levels of obesity may be linked as much to moral beliefs around ‘normality’ and weight, as they are to actual health risks. But seldom are the public invited to explore the ways in which the sort of moral panic created by the obesity discourse may be damaging to people’s health through shame-based narratives it endorses around the body, eating and weight” (Rich and Evans, 2005, pg. 344). Health-related fatphobia, while often verbalized out of professed pity or concern for fat people’s well-being, ulitmately reinforces anti-fat stigma; it is yet another means for excluding that which is seen as abnormal and enforcing the moral superiority of the powerful elite.

Similarly, the fat body is seen as a sign of deviance from signal American values of individualism, self-discipline, and exceptionalism, which are associated with thinness; conversely, fatness conveys ignorance, self-abandonment, and ultimately poverty (Townend, 2009, pg. 181). Under the influence of fatphobia, we ascribe fatness to personal moral failure and individual lifestyle choice rather than uncontrollable factors such as genetics, socioeconomic status, and access to healthy food. As Crandall (1994) confirms, this perception of controllability significantly increases the moralization of body weight: “To generate dislike of fat people, one must think fat undesirable and simultaneously blame the person for [their] situation. [...] Fat people appear to be just one more on a long list of deviant groups.” The alignment of fatness with an aberrant or norm-defying character has an othering and subordinating effect; it leads to anti-fat discrimination in academic, professional, and social settings, where fat people are seen as incompetent, lazy, undisciplined, uncivilized, and immoral (Manne, 2024). These biases lead to lower salaries, disadvantages in university admissions and employment, and weight-based harassment.

Weight bias denies fat people equitable access to education, healthcare, comfortable clothing and seating, and career opportunities on the cultural level; beyond this, it leads them to be body-shamed, blamed for having excessive weight, and treated as burdens through overtly discriminatory language. The projection of morality onto body weight deepens anti-fat and pro-thin bias and serves to rationalize the continued oppression of fat bodies (Saguy, 2012; Saguy and Gruys, 2010; Manne, 2024). Considering this, alongside the fact that moralization predicts exacerbated anti-fat stigma and greater stigma internalization on the part of fat people (Ringel and Ditto, 2019; Täuber et al., 2018), we find it critical to investigate weight bias in terms of morally valenced language.