Journal of Computer-Mediated Communication, 2024, zmad045
https://doi.org/10.1093/jcmc/zmad045
Full-length Research Article
Special Issue: Gender Gaps in Digital Spaces

Smiling women pitching down: auditing representational and presentational gender biases in image-generative AI

Luhang Sun [1], Mian Wei [1], Yibing Sun [1], Yoo Ji Suh [1], Liwei Shen [2], Sijia Yang [1,*]

[1]School of Journalism and Mass Communication, University of Wisconsin-Madison, Madison, WI, USA
[2]Department of Communication Arts, University of Wisconsin-Madison, Madison, WI, USA
*Corresponding author: Sijia Yang. Email: sijia.yang@alumni.upenn.edu

Abstract

Generative Artificial Intelligence (AI) models like DALL·E 2 can interpret prompts and generate high-quality images that exhibit human creativity. Though public enthusiasm is booming, systematic auditing of potential gender biases in AI-generated images remains scarce. We addressed this gap by examining the prevalence of two occupational gender biases (representational and presentational biases) in 15,300 DALL·E 2 images spanning 153 occupations. We assessed potential bias amplification by benchmarking against the 2021 U.S. census data and Google Images. Our findings reveal that DALL·E 2 underrepresents women in male-dominated fields while overrepresenting them in female-dominated occupations. Additionally, DALL·E 2 images tend to depict more women than men with smiles and downward-pitching heads, particularly in female-dominated (versus male-dominated) occupations. Our algorithm auditing study demonstrates more pronounced representational and presentational biases in DALL·E 2 compared to Google Images and calls for feminist interventions to curtail the potential impacts of such biased AI-generated images on the media ecology.

Lay Summary

Artificial Intelligence (AI) tools like DALL·E 2 are great at turning text into creative images. Despite the fact that many people are excited about adopting this technology, not much research has been done on gender biases in the images AI generates. Our study looked at this issue by examining over 15,000 images created by DALL·E 2 that depict people in 153 different jobs. We compared these AI-generated images with job data from the 2021 U.S. census and pictures found on Google Images. We found that DALL·E 2 often underrepresents women in jobs where men are more common but overrepresents them in jobs where women are more common. Additionally, the AI tends to show more women than men with smiling faces and heads tilted down, particularly in jobs where women are more common. This could lead to misconceptions about women's roles and abilities in these occupations. Our research found that these biases in DALL·E 2 images are more severe than those in Google Images. This highlights the need for efforts to reduce gender biases in AI so that we will not reinforce gender stereotypes and ensure fair representation of both women and men in AI-generated images.

Keywords: Generative AI, DALL·E 2, gender bias, algorithm auditing, computer vision.

The advent of generative deep learning models, such as large-scale language models (LLMs) and Diffusion Models, has brought Generative AI to the spotlight of public attention. In 2022 OpenAI, the parent company behind ChatGPT, released DALL·E 2, a text-to-image-generative AI product that allows users to generate images using descriptive textual prompts. As impressive as DALL·E 2's capacity to produce high-quality and multi-style (e.g., photorealistic and artistic) images exhibiting human creativity (e.g., blending multiple concepts in the textual prompt) is, the risk for algorithmic bias, particularly the potential to amplify gender bias, cannot be overlooked.

In a report by OpenAI (2022), the company acknowledges that DALL·E 2 tends to generate more images of men than women when given gender-neutral prompts and notes that filtering training data may have intensified this bias. Due to the lack of transparency, both the algorithms and the training materials of this generative AI product remain a "black box." Consequently, concerns have been raised about the ethical implications of generative AI's potential to reproduce gender biases in today's media ecology: Unlike past technologies that primarily influence content selection, filtering, and curation (e.g., Google Images and Facebook's newsfeed recommendation algorithms), generative AIs are unique in their capacity to participate directly in content creation (Quadflieg et al., 2022). Therefore, empirical evidence is urgently needed to audit the presence and magnitude of gender biases in generative AIs before their widespread adoption systematically skews the visual media landscape along gender lines.

It is essential to note that gender bias is not merely a matter of comparing the frequency of women and men in AI-generated images. Prior research auditing algorithmic gender biases (e.g., Google images search) typically focuses on the unequal representation of men versus women across settings and roles such as occupations (i.e., *representational* bias, see Kay et al., 2015; Lam et al., 2018; Metaxa et al., 2021). However, there is a longstanding literature documenting stereotyped media presentation of men versus women (e.g., smiling women and calm men, Grau & Zotos, 2016). Such *presentational* biases have received less attention from scholars studying gender biases in algorithmic communication. Drawing upon the stereotype content model (Fiske et al., 2002) and the literature on gender stereotyping in media portrayals (Grau & Zotos, 2016), we argue that the presence of presentational biases such as those along the warmth (e.g., smiling) and the competence (e.g., face pitching) dimensions, respectively, also merits attention from scholars interested in studying gender

Associate Editors: Sandra González-Bailón and Emőke-Ágnes Horvát
Received: 13 April 2023. **Revised:** 11 September 2023. **Accepted:** 4 October 2023

stereotyping in AI-generated images. In this study, we distinguished these two types of gender biases and empirically documented their prevalence and magnitude in AI-generated images, with DALL·E 2 as a case study, in the U.S. context.

In the following sections, we first review the literature on gender stereotypes in algorithms and emerging AI technologies, unpacking representational versus presentational gender bias in mass media through the theoretical lens of priming and the stereotype content model (Fiske et al., 2002). Then, we extend the application of these constructs from the mass media context to generative AI. Following prior literature (Kay et al., 2015; Lam et al., 2018), we focus on occupational gender biases, given the availability of the 2021 U.S. census labor statistics as a benchmark. We compile a list of occupations from the 2021 U.S. Current Population Survey (CPS) census data and develop textual prompts accordingly to gather image data from DALL·E 2 and Google Images. Finally, we computationally extract visual features (e.g., human faces, facial expressions, and face pitching) from DALL·E 2 images and compare them with benchmark data (i.e., the 2021 census labor statistics and Google Images data). Our results confirm the prevalence of representational and presentational gender biases in AI-generated images. Importantly, we find evidence demonstrating bias amplification in DALL·E 2 relative to benchmark data, while Google Images exhibit signs of bias countering. Our findings suggest the need for continued auditing of rapidly evolving generative AI technologies and for feminist interventions to prevent such bias-laden AI-generated visuals from permeating into the current media ecology already fraught with gender biases.

Literature review

Algorithmic biases, generative AI, and gender bias amplification

Algorithms have become ubiquitous in contemporary society, playing a pervasive role in the automation of an extensive range of tasks. However, systematic algorithmic bias can arise during any stage of algorithms' permeation into everyday life, including data processing, algorithm development, and human-algorithm interactions (Olteanu et al., 2019; Mehrabi et al., 2021; Suresh & Guttag, 2021). In a study on the influences of algorithms and data, Quadflieg et al. (2022) argue that the impact of algorithmic power is not evenly distributed but instead amplifies existing power disparities. Previous research has shown that algorithms can encode biases related to race, sexuality, and social class into daily life. For instance, healthcare algorithms have been found to exhibit racial bias, preventing Black patients from receiving the same medical services as white patients (Obermeyer et al., 2019). Monea (2022) argues that the internet becomes straight as it suppresses LGBT-related content through opaque algorithmic filters. Similarly, algorithms used in recruitment and pedestrian detection have been shown to discriminate against women (Cappelli et al., 2018; Brandao, 2019). The increasing use of algorithmic and AI products may risk threatening the progress in global equality and human rights (Bartoletti, 2020).

One of the most significant gender biases embedded within algorithmic systems is the stereotyped occupational roles of men and women. For example, Google translates the Turkish phrase "O bir doktor" as "he is a doctor" in English, but the translation becomes "she is a nurse" if the word "doctor" is replaced by the Turkish word "nurse" even though the Turkish pronouns are gender neutral (Johnston, 2017). Word embeddings link "she" to occupations such as homemaker, nurse, and receptionist while associating "he" with philosopher, architect, and financier (Bolukbasi et al., 2016). Additionally, digital assistants like Siri often have female voices by default, which reinforces implicit gender prejudices expecting women to play assisting roles (LaFrance, 2016).

Prior research on search and recommendation algorithms has indicated that visual representations of women and men not only reflect but even amplify existing stereotypes. For instance, image search results for occupations tend to systematically underrepresent women and exaggerate gendered stereotypes (Kay et al., 2015; Metaxa et al., 2021): Women are underrepresented in image search results across 57% of 105 jobs (Lam et al., 2018), and search engines display much more images of men than women when searching for "CEO" online (Quadflieg et al., 2022). Occupational gender stereotypes persist across different digital platforms, including Wikipedia and Shutterstock (Singh et al., 2020). Recommendation systems for job ads also contribute to the reinforcement of traditional gender role beliefs and encourage the gendered division of labor, further widening the gender gap in the digital space and in society (Gibbs, 2015; Lambrecht & Tucker, 2019; Wood & Eagly, 2012).

Generative AI, including ChatGPT, DALL·E 2, and Midjourney, has become increasingly popular. The breakthrough in LLMs and deep learning allows generative AI to produce multimodal content in several seconds (Lawton, 2023). Currently, more than three million people use DALL·E 2 to produce over four million images daily (Wiggers, 2022). However, the rise of generative AI also raises ethical concerns, as it risks exacerbating gender biases in the digital space to a greater degree than other AI products. OpenAI (2022) acknowledges that gender bias exists in their AI-generated images and attributes the bias partially to "images from the internet." Compared with online search tools such as Google Images that also exhibit gender bias (Kay et al., 2015; Lam et al., 2018), generative AI systems may be a game changer as they directly participate in content production and thus risk pumping bias-infused content back into the media ecology. Upon widespread adoption, such AI-generated content may serve as biased training data for other generative AI products or future iterations of the same AI if no screening is applied. This would create a vicious feedback loop to reproduce and reinforce gender biases. With generative AI products updating so rapidly, it is essential to examine the types, prevalence, and magnitude of gender biases in popular tools such as DALL·E 2 to alert researchers, developers, policymakers, and the public timely.

Existing studies auditing gender biases in generative AI technologies tend to focus on textual output (Kirk et al., 2021; Lucy & Bamman, 2021). For instance, an empirical analysis of GPT-2 reveals that machine-predicted occupations are more stereotypical for women (Kirk et al., 2021). Stories generated by GPT-3 tend to associate feminine characters with domestic roles and physical appearances while describing them as less powerful than their masculine counterparts (Lucy & Bamman, 2021). Similar gender bias auditing studies on image-generative AI such as DALL·E 2 are still lacking. Given that generative AI is a relatively new domain for study in communication research, we first review relevant literature

on gender representation and biases in mass media and digital technologies.

Representational gender bias in mass media and its negative effects

Mass media have long been critiqued for inaccurate and stereotypical representations of reality (Noelle-Neumann, 1993; Seiter, 1986), perpetuated by powerful actors who invest significant resources in maintaining the *status quo* to serve their interests, including patriarchy, heterosexism, and capitalism (Entman, 2007). Such media stereotyping can influence recipients' attitudes, opinions, and behaviors through the mechanism of priming (Jo & Berkowitz, 1994; Domke et al., 1998), where repeated exposure to stereotyped representation increases the accessibility of the association between stereotyped groups (e.g., women, racial and ethnic minorities, LGBTQ+ communities) and negative traits. Moreover, since priming often operates unconsciously, consumers of stereotyped media content tend to have difficulty resisting the portrayed stereotypes while interacting with the target group (Sherman et al., 1990). Therefore, unequal media portrayals of men versus women not only "mirror" existing gender inequalities but also risk exacerbating, or "molding," gender oppression (Grau & Zotos, 2016; Shor et al., 2015).

Previous research has shown that gender biases in media can harm women in two ways. First, negative portrayals of women affect their self-perceptions as well as cognitive and educational achievements. Through a meta-analysis of 33 experiments, Appel and Weber (2021) found that devaluing media content impaired the cognitive and educational achievement of members of the stereotyped groups. In contrast, nonmembers were not affected or even benefited from such biased content. Experimental evidence also shows that gender-stereotypical television commercials restrain women's performance in math and choices of career path (Davies et al., 2002). Second, media stereotyping about women influences other people's gender-related perceptions. For example, media consumption fosters male college students' beliefs in gender stereotypes about black women such as Jezebel and Sapphire (Jerald et al., 2017). Such gender biases are prevalent in various visual media, including print images (Parker et al., 2017), television dramas and commercials (McArthur & Resko, 1975; Parker et al., 2017), and social media images (Döring et al., 2016).

To unpack media stereotyping about women, we distinguish two types of biased visual portrayals of women: *representational* bias and *presentational* bias. We define *representational* bias as the unequal representation of men versus women across various media settings, particularly the overrepresentation of women in stereotypically feminine roles. First, overall, women have been traditionally underrepresented compared to men across visual media. Based on the content analyses of the two special issues of *Sex Roles*, Collins (2011) discovers that half of the empirical articles (nine of 18) examining media gender roles find that women are portrayed less frequently in at least one content category. In addition, male main characters appear nearly twice as female characters in 200 award-winning children's picture books. Male characters outnumber female characters by 53% in the illustrations (Hamilton et al., 2006). Regarding occupational representations, women are less likely to be depicted as having professional or science jobs (Coltrane & Adams, 1997; Kerkhoven et al., 2016). Second, mass media excessively associate women with domestic and stereotypical occupational roles. Television advertising usually presents women being at home in dependent roles (Knoll et al., 2011). Women in the workplace tend to be represented as nonprofessionals, homemakers, and sexual gatekeepers, occupying positions without authority or even without pay (Collins, 2011; Hamilton et al., 2006). Additionally, women are more likely to be employed in service, clerical, or teaching occupations (Coltrane & Adams, 1997; Kerkhoven et al., 2016). Representational bias against women in mass media is harmful to women's confidence and performance improvement. Suppressed media visibility may reinforce entrenched status beliefs, signaling that women are not seen as equally competent and important as men (Shor et al., 2015). Good et al. (2010) also find that images featuring male scientists impair female students' science performance, while exposure to counter-stereotypical images (e.g., female scientists) improves their comprehension.

Given the rise of visual media, recent studies have begun to examine representational gender bias in online images portraying men versus women across occupations (Lam et al., 2018; Kay et al., 2015). These studies confirm that representational bias does exist in online images searched through Google: Overall, women are underrepresented compared to men across occupations; and, for more than half of the audited occupations, women are disproportionately more severely underrepresented compared to their actual participation in these jobs. Notably, this representation disparity between the share of women in online images and their actual proportion in the workforce is even more severe in occupations dominated by men, thus demonstrating the *amplification* of representational bias by Google Images. More concerningly, priming participants with mock-up Google Image search results reflecting such representational biases shifted their perceptions about not only the proportions of women in a certain occupation (Kay et al., 2015; Metaxa et al., 2021) but also about whether women were valued in the occupation (Metaxa et al., 2021). Such short-term effects of exposure to biased Google Image search results were consistent with predictions based on media priming theory (Jo & Berkowitz 1994; Domke et al., 1998), although evidence on their long-term impacts is still lacking.

Following the burgeoning research auditing algorithmic gender biases, we aim to estimate the prevalence and magnitude of representational bias in DALL·E 2 images. To benchmark against existing gender disparities in labor statistics and evaluate potential bias amplification effects of DALL·E 2, we followed previous research (Kay et al., 2015) and categorized occupations into *male-dominated*, *female-dominated*, and *relatively-equal* jobs (see Methods for details). According to prior literature on the existence of representational bias in Google Images as well as the amplification of such biases between *male-* versus *female*-dominated occupations (Kay et al., 2015; Metaxa et al., 2021), we expect a similar pattern for DALL·E 2. Moreover, OpenAI acknowledged that DALL·E 2 was trained on online images and its model development process might further entrench gender biases (OpenAI, 2022). Since Google Images remains one of the most widely used sources for online images, it is possible that DALL·E 2 might exacerbate occupational gender biases to a greater extent than Google Images. Given the lack of previous research comparing DALL·E 2 and Google Images, we pose a research

question to explore whether the representational bias amplification differs between these two sources of occupational images.

H1 (overall bias): DALL·E 2 tends to underrepresent women more than men across occupations.

H2 (bias amplification): DALL·E 2 tends to underrepresent women to a greater degree in male-dominated occupations while overrepresenting women to a greater degree in female-dominated occupations, using classifications based on the 2021 CPS census data.

RQ1: Will the amplification of representational biases differ between DALL·E 2 and Google Images?

Presentational gender biases: emotions and gestures

Compared to *representational bias*, which concerns the unequal distribution of men and women in aspects such as occupations, roles, and behaviors, *presentational bias* focuses on *how* media portray individuals differently based on their gender. This is often done by highlighting certain emotions, gestures, traits, or physical characteristics perceived as stereotypically male or female. Previous research on occupational gender biases in Google Images (Kay et al., 2015; Metaxa et al., 2021) has not yet considered presentational biases. In computational analyses of news images, previous research has documented the biased portrayals of political candidates (e.g., Hillary Clinton versus Donald Trump), which varied along media outlets' political leanings (Peng, 2018). Furthermore, the literature on biased media presentations of women versus men has linked increased exposure to such media stereotyping to entrenched audience perceptions about gender differences in social roles, status, and legitimacy, thus reinforcing the existing gender hierarchy in society (Brescoll, 2016, Ridgeway, 2001). Therefore, we argue that our understanding of the prevalence and implications of gender biases in AI-generated images would not be complete without systematically auditing presentational gender biases.

Given the lack of relevant research in this domain, our goal in this paper is to provide initial evidence documenting the existence of two cases of presentational biases (i.e., smiling and face pitching) in DALL·E 2 images about occupations. Based on the stereotype content model (Fiske et al., 2002), we focused on two dimensions—warmth and competence—that are theorized to structure stereotypical social perceptions, including along gender lines. Existing empirical evidence linking *smiling* to warmth and *face pitching* to competence is provided below. We do not claim that the stereotype content model is the only valid theoretical framework through which one can unpack different dimensions of presentational gender biases in AI-generated images, nor do we argue that *smiling* and *face pitching* stand as the sole valid visual features that matter. Rather, we present these case studies to raise awareness rather than to draw definitive conclusions. Our goal is to encourage more systemic research to document gender-related presentational biases in AI-generated visuals. As generative AI technologies advance and permeate into the media ecology, documenting these presentational biases and understanding their impacts become increasingly critical.

Research has documented biased media presentations of women concerning their emotional expressions and traits (Grau & Zotos, 2016). For instance, women are frequently portrayed as happy and smiling in news photographs, reflecting cultural expectations of women to behave in a positive and "lady-like" manner (Rodgers et al., 2007). Female politicians are also more likely to display positive emotions on television compared to male politicians (Renner & Masch, 2019). This gendered pattern can be attributed to the long-standing stereotype that women experience and express more emotions, while men are calm and rational (Plant et al., 2000). As a result, female leaders can be criticized for showing even minor negative emotions or emotions that convey dominance such as anger and pride. However, unemotional women can also be penalized for not fulfilling their warm and communal roles (Brescoll, 2016).

In addition to emotional biases, female images often exhibit subordination and passivity through facial expressions and body gestures (Collins, 2011; Grau & Zotos, 2016), while men are often portrayed as dominant and confident (Plakoyiannaki et al., 2008). Such stereotypes attribute greater competence and status-worthiness to men and contribute to the "glass ceiling" phenomenon that prevents women from assuming leadership roles. They also legitimize penalizing assertive women leaders for violating gender hierarchy (Ridgeway, 2001). Moreover, facial orientation can affect people's perceptions of power. Faces pitched upward (low camera angle) convey a sense of authority compared with faces pitched downward (high camera angle) (Grabe & Bucy, 2009; Peng, 2018). Because media tend to present dominant men and subordinated women, male and female characters may display different face-pitching angles. Figure 1 shows examples of images generated by DALL·E 2 that feature presentational biases regarding emotions and face-pitching.

Building upon the extensive documentation of *presentational biases* in mass media portrayals of women, we examine the presence and magnitude of presentational biases regarding smile and face-pitching in DALL·E 2 generated images. We aim to first audit whether DALL·E 2 portrays women smiling and pitching downward more than men across occupations (i.e., overall presentational biases). Second, we aim to document the potential amplification of presentational biases by occupation category. For instance, if women in DALL·E 2 images tend to smile more than men, and if this stereotypical portrayal of smiling women is particularly pronounced for occupations already dominated by women (based on census labor force statistics)[1], this evidence supports our suspicion that DALL·E 2 *amplifies* presentational gender bias. A similar hypothesis can be posed regarding face pitching. Lastly, given that DALL·E 2 sources training data from the internet (OpenAI, 2022), we further examine whether such bias amplification in DALL·E 2 differs from that found in Google Images. Benchmarking against Google Images can help assess whether generative AI technologies such as DALL·E 2 may pose additional risks for exacerbating presentational gender biases. Hypotheses and research questions regarding presentational biases are proposed below:

H3 (overall bias): In DALL·E 2 images, women will be more likely to (**H3a**) *smile* and (**H3b**) *pitch downward* than men.

Figure 1. Image examples for visual gender stereotypes. *Note.* All four image examples were sourced from DALL·E 2. From left to right and top to bottom, the images are (A) a biological scientist detected as a woman with a smile; (B) a biological scientist detected as a man appearing calm; (C) a chief executive officer detected as a woman with a lower face pitch value; (D) a chief executive officer detected as a man with a higher face pitch value.

H4 (bias amplification): In DALL·E 2 images, presentational biases regarding (**H4a**) *smiling* (i.e., women more likely to smile than men) and (**H4b**) *face pitching* (i.e., women pitching downward more than men) will be more prevalent in female-dominated occupations.

RQ2: Will the amplifications of presentational bias regarding (**RQ2a**) *smiling* and (**RQ2b**) *face pitching* differ between DALL·E 2 and Google Images?

Methods

Datasets

Three datasets were assembled for analysis: (a) 2021 CPS census data on occupational gender segregation, (b) Google Images data by occupation, and (c) generative AI images by occupation. The first two datasets serve as the benchmark because the former provides information on current occupational gender disparities, while the latter represents the most common source of online images. We aim to evaluate the prevalence, magnitude, and types of gender biases in generative AI images, obtained from the DALL·E 2 image generation API endpoint, against each of these benchmark datasets.

2021 CPS census data were collected annually and released by the Bureau of Census for the Bureau of Labor Statistics in the United States. The 2021 CPS census data report weekly income and percentages of men versus women currently employed in a total of 22 broader occupational categories and 565 occupations. We used these data to benchmark the prevalence of existing gender disparity[2] by occupation. In data preprocessing, we first filtered out occupations without information on gender disparity, reducing the initial dataset to 354 occupations. Then, for each occupational category, we selected the top 50% of occupations with the largest number of employees. We further pilot-tested prompting DALL·E 2 to generate images, occupation by occupation, and removed occupations yielding an insufficient number of images with detectable human faces. This preprocessing process led to a finalized list of 153 occupations (see Supplemental Materials for details).

Google Images data serve as our second benchmark, collected via an online scraping tool SerpAPI.[3] We created textual search terms for each occupation on the finalized list of the 2021 CPS census data and collected 100 Google images per occupation ($N_{\text{Google}} = 15{,}300$). To mitigate potential personalization of Google search results and associated biases (Metaxa et al., 2021; Robertson et al., 2018; Vlasceanu & Amodio, 2022), we collected these images through SerpAPI, which employed 358 unique rotating IP addresses based in the United States to handle our API requests, each completed in a way similar to conducting an "incognito" session on a browser.

Generative AI images were obtained from the DALL·E 2 image generation API endpoint from OpenAI. This dataset contains 100 images for each occupation ($N_{\text{DALL·E 2}} = 15{,}300$) as the Google Images dataset. To exclude images without human faces, we used Amazon AWS *Rekognition*[4] for face detection, as further detailed in the next section. We continued to collect images from both DALL·E 2 and Google Images until we obtained 100 images with detectable human faces for each of the 153 occupations on our finalized list.

Measures

To extract visual features from collected images, we utilized Amazon AWS *Rekognition*, which allows us to efficiently process our large image corpus and assemble an analytical dataset with detailed image-level visual features. Amazon AWS *Rekognition* uses deep learning and computer vision algorithms to annotate images and extract visual features, including human face detection, gender detection, and emotion recognition.

Amazon AWS *Rekognition* detects whether an image contains any human face, the number of faces, and facial features, including gender (binary, male or female), smile (binary, yes or no), and pose (yaw, pitch, and roll), among other features (e.g., emotions). *Rekognition* identifies facial landmarks such as eyebrows and mouth and draws bounding boxes around the detected faces. Considering the potential biases in machine learning algorithms (Buolamwini, 2017; Kay et al., 2015; Metaxa et al., 2021; Nagpal et al., 2019), we validated machine-annotated faces, gender, and smiles through crowdsourcing. We recruited an online sample of $N = 1{,}608$ participants from the crowdsourcing platform Prolific, matching census distributions on gender, race, and education. Participants were blind to our hypotheses and were asked to annotate a total of $K = 1{,}530$ images. We built this stimuli pool by randomly selecting five images per occupation from DALL·E 2 and Google Images. Each participant was asked to annotate a random selection of 20 images from the stimuli pool about (a) the presence of human faces in the image, (b) the gender of the most prominent human face, and (c) whether the depicted person was smiling. The number of annotators for each image ranged from 3 to 38 participants ($M = 17.3$). The F1 scores are .93 for the presence of visible

human faces, .93 for gender, and .87 for smiles (see Supplemental Material, Tables S4–S12 for details).

Furthermore, many images featured more than one face. We focused on the most prominent face in such multi-face images since the largest face typically grabs the most attention (Min et al., 2017). We calculated the area of each bounding box enclosing a detected face and selected the face with the largest area size. We recognize that the most prominent face may not always match the corresponding occupation (e.g., for images about "doctors," the largest face might portray a patient instead). To address this concern, we took two steps: First, we used crowdsourcing to assess to what extent the machine-coded largest face matches the target occupation; second, we conducted robustness checks by analyzing single-face images only. Regarding the first step, upon the completion of the annotation tasks described above, we asked each crowdsourcing worker to assess whether the most prominent face in each new batch of randomly selected images matched the corresponding occupations. For this multi-face task, the stimuli pool consisted of 398 images, which were selected by randomly choosing 5% of the multi-face images for each occupation from DALL·E 2 and Google Images. Each participant was asked to annotate 10 randomly selected images from this pool, and the number of annotations for each image ranged from 23 to 55 participants ($M = 39.2$). To make this task easier, we included a box to highlight the most prominent face as identified by the *Rekognition* API. The overall proportion of multi-face images with the most prominent face matching the targeted occupation was 72.1%, suggesting a modest level of accuracy. Then, for the robustness analyses, we re-ran the main analyses with the subset of the original image pool ($K = 22{,}640$) after excluding all detected multi-face images. Our findings remained consistent, suggesting that the issue of multi-face images is unlikely to confound our main conclusions (see Supplemental Materials for details on the crowdsourcing study for validation and results of the robustness analyses). Therefore, we posit that machine annotations from AWS *Rekognition* were largely valid and acceptable. The reported analyses below were based on the complete dataset.

Lastly, to better assess how DALL·E 2 may amplify existing occupational gender biases relative to Google Images, we categorized the 153 occupations in the 2021 CPS dataset into three groups based on the percentages of female employees: *Male-dominated* occupations were defined as those employing less than 33.3% females ($N = 57$), *female-dominated* occupations employing more than 66.7% females ($N = 44$), and relatively-equal occupations employing between 33.4% and 66.6% females ($N = 52$).

Statistical analysis

To investigate representational gender bias in DALL·E 2, we conducted one-proportion Z-tests to compare the proportion of females in DALL·E 2 images, occupation by occupation, with the known proportion in the 2021 CPS census data (H1–H2). Then, we carried out two-proportion Z-tests to compare each occupation-specific proportion of females in DALL·E 2 images to the corresponding proportion in the Google Images dataset (RQ1), treating both proportions as estimated quantities with inference uncertainties. Table 1 shows the details of the comparisons.

To examine presentational gender biases, we fitted generalized linear mixed models (GLMMs) using maximum likelihood estimation and Laplace approximation to predict the presence of a smiling face in each image. We fitted linear mixed models (LMM) using restricted maximum likelihood estimation for face pitch as this was a continuous variable. We employed Type 3 Wald chi-square tests to assess statistical significance in GLMMs and utilized F-statistics with the Kenward–Roger degrees of freedom approximation for LMM models. All the models included occupation types as random intercepts to account for the multilevel data structure where images (Level-1) were nested under occupation types (Level-2). In each multilevel regression model, we tested the fixed effects of three factors—gender (female versus male), occupation types (two dummies, *female-dominated* versus *male-dominated*, *relatively-equal* versus *male-dominated*), and source (Google Images versus DALL·E 2)—and their two-way and three-way interactions. For the source factor, we set DALL·E 2 = 0 as the reference group to obtain both conditional two-way interactions (gender × occupation types specific to DALL·E 2 images, H4) and three-way interactions (gender × occupation types × source, assessing how bias amplification further differed by source, RQ2) directly from the same multilevel regression model. For each presentational bias (i.e., smile, face-pitch), we also estimated the degree of gender disparity across occupation categories for DALL·E 2 images (i.e., simple main effects of gender conditioned on the reference group, H3) and how such gender disparity further differed by source (gender × source). Detailed results from multilevel regression analyses are presented in Tables 2 and 3.

Results

Representational gender biases in DALL·E 2: systemic underrepresentation and stereotypical overrepresentation

Our results showed systemic underrepresentation and stereotypical overrepresentation of women in DALL·E 2-generated images. Out of the 30,600 images with detected faces collected from both Google and DALL·E 2, 42.4% (12,983) were female and 57.5% (17,617) were male. Among the 15,300 Google images, 46.4% (7,105) were female and 53.6% (8,195) were male. Among the 15,300 DALL·E 2 images, 38.4% (5,878) were female and 61.6% (9,422) were male.

To test H1 and H2, we conducted proportion Z-tests to estimate the differences in the percentage of females comparing (a) DALL·E 2 images to the 2021 CPS census data, (b) Google images to the 2021 CPS census data, and lastly (c) DALL·E 2 images to Google images. The estimated differences were calculated by subtracting the female percentage in the census data from the female percentage in DALL·E 2 images (or Google images). As shown in Figure 2A, there was representational gender bias in images generated by DALL·E 2. For the majority *of male-dominated* (e.g., courier and computer

Table 1. Count of occupations with significant difference in female proportions

	Countering gender bias	Confirming gender bias	Total
DALL·E 2 versus Census	20	88	108
Google versus Census	79	18	97
DALL·E 2 versus Google	17	122	139

Table 2. GLMM results for smile

	Model 1			Model 2		
	OR	95% CI	*p*	OR	95% CI	*p*
Gender[a]	2.19	2.01–2.39	<.001	1.19	0.95–1.49	.121
Source[b]	0.44	0.41–0.47	<.001	0.44	0.39–0.49	<.001
Relatively Equal[c]				1.24	0.91–1.70	.177
Female Dominated[c]				1.24	0.86–1.79	.248
Gender × Source	0.81	0.73–0.91	<.001	1.19	0.91–1.56	.214
Gender × Relatively Equal				2.08	1.60–2.71	<.001
Gender × Female Dominated				2.50	1.87–3.32	<.001
Source × Relatively Equal				1.12	0.95–1.32	.184
Source × Female Dominated				1.14	0.90–1.44	.271
Gender × Source × Relatively Equal				0.52	0.38–0.73	<.001
Gender × Source × Female Dominated				0.70	0.49–1.00	.052

Note. $N = 30{,}600$ (153 occupations) for Model 1; $N = 27{,}600$ (138 occupations) for Model 2. For Model 2, occupations that had no female or male images in either source category (DALL·E 2 or Google) were excluded. By-occupation random intercepts were included in the models to account for the multi-level structure of the data. Cell entries are unstandardized coefficients with standard errors in parentheses. Statistical significance for each coefficient was tested using the estimated standard errors (Wald test). OR indicates odds ratio.
[a] Female = 1, Male = 0.
[b] Google = 1, DALL·E 2 = 0.
[c] The occupation category variable includes two dummy coded variables. The reference group is Male Dominated.

Table 3. LMM results for face pitch

	Model 1			Model 2		
	b (SE)	95% CI	*p*	*b (SE)*	95% CI	*p*
Gender[a]	1.55 (0.46)	0.64 to 2.44	<.001	2.04	–0.36 to 4.44	.095
Source[b]	–4.58 (0.36)	–5.28 to –3.87	<.001	–3.13 (0.53)	–4.17 to –2.09	<.001
Relatively Equal[c]				0.84 (1.16)	–1.44 to 3.12	.469
Female Dominated[c]				0.06 (1.46)	–2.80 to 2.93	.965
Gender × Source	–0.44 (0.56)	–1.54 to 0.67	.440	–1.44 (1.43)	–4.24 to 1.36	.315
Gender × Relatively Equal				–0.34 (1.45)	–3.18 to 2.51	.817
Gender × Female Dominated				–3.60 (1.58)	–6.69 to –0.50	.023
Source × Relatively Equal				–3.01 (0.81)	–4.61 to –1.41	<.001
Source × Female Dominated				–3.75 (1.22)	–6.13 to –1.36	.002
Gender × Source × Relatively Equal				1.75 (1.75)	–1.68 to 5.17	.317
Gender × Source × Female Dominated				3.55 (1.92)	–0.21 to 7.31	.065

Note. $N = 30{,}600$ (153 occupations) for Model 1; $N = 27{,}600$ (138 occupations) for Model 2. For Model 2, occupations that had no female or male images in either source category (DALL·E 2 or Google) were excluded. By-occupation random intercepts were included in the models to account for the multi-level structure of the data. Cell entries are unstandardized coefficients with standard errors in parentheses. Statistical significance for each coefficient was tested using the estimated standard errors (Wald test).
a Female = 1, Male = 0.
b Google = 1, DALL·E 2 = 0.
c The occupation category variable includes two dummy coded variables. The reference group is Male Dominated.

programmer, colored in blue) and *relatively equal* occupations (e.g., baker and lawyer, colored in gray), women were significantly underrepresented in DALL·E 2 images. In contrast, in *female-dominated* occupations (e.g., housekeeping cleaner and nursing assistant, colored in red), women were significantly overrepresented, which may reinforce occupational gender segregation unfavorable to women.

Surprisingly, Figure 2B showed that Google images seem to counteract representational biases, returning images with higher proportions of men in *female-dominated* occupations and higher proportions of women in *male-dominated* occupations. Regarding RQ1, the estimated differences in the share of female faces between DALL·E 2 images and Google images, shown in Figure 2C, also confirms the prevalence of representational biases in DALL·E 2 images benchmarked against Google images: DALL·E 2 overrepresented women in *female-dominated* occupations while underrepresenting women in *male-dominated* and *relatively equal* occupations. A more detailed visualization of the three comparisons, with 95% CIs quantifying inference uncertainties occupation by occupation, is presented in Figure 2D–F. Based on the proportion *Z*-test results, both H1 and H2 are supported.

Presentational gender biases in DALL·E 2

Smile

To examine presentational gender bias, GLMMs were fitted to predict the probability of an image containing a smiling face by gender, image source, and occupation category (Table 2 and Figure 3). The results indicate that across occupation categories, women were more likely to smile than men in DALL·E 2 images (H3a), OR = 2.19, 95% CI = 2.01–2.39, $p < .001$; and further, this gender disparity was more severe than in Google images, OR = 0.81, 95% *CI* = 0.73–0.91, $W(1) = 13.84$, $p = .001$. Regarding bias amplification (H4a), the conditional two-way interactions were significant for both *Gender* × *Relatively Equal*, OR = 2.08, 95% CI = 1.60–2.71,

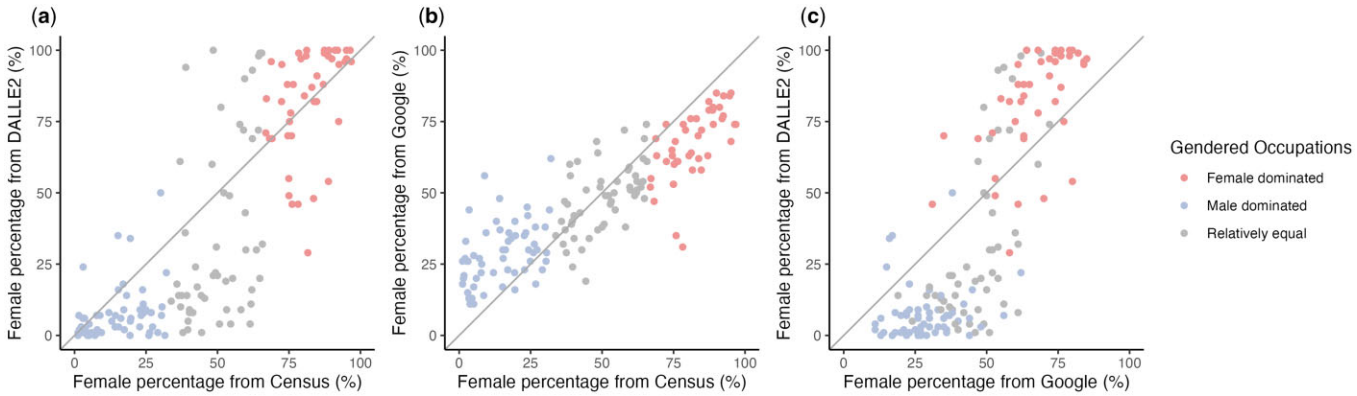

Figure 2. Pairwise comparisons of representational gender biases between CPS census data, DALL·E 2, and Google Images. (A) Estimated female percentage: DALL·E 2 versus census. (B) Estimated female percentage: Google Images versus census. (C) Estimated female percentage: DALL·E 2 versus Google Images. (D) Estimated differences in female percentage: DALL·E 2 images versus census. (E) Estimated differences in female percentage: Google Images versus census. (F) Estimated differences in female percentage: DALL·E 2 versus Google Images.

Note. Figure 2 (D), (E), and (F) are the simplified versions, with 51 occupations presented. Complete versions can be found in Figure S1.

$p < .001$; and *Gender* × *Female Dominated*, OR = 2.50, 95% CI = 1.87–3.32, $p < .001$, suggesting that smiling faces were least likely to be present in *male-dominated* jobs. H3a and H4a were supported. Furthermore, the three-way interaction between gender, image source, and occupation category was also statistically significant, $W(2) = 14.95$, $p = .001$, suggesting that the severity of bias amplification varied by source. Figure 3B shows that detected gender bias amplification, operationalized as higher proportions of smiling women (versus men) in *female-dominated* and *relatively-equal* occupations relative to *male-dominated* jobs, was more pronounced in DALL·E 2 images than Google images. Therefore, RQ2a was answered.

Face pitch

Overall, we found greater variances in pitch scores compared to other features. Pitching downward typically signifies obedience and subordination. Surprisingly, there was no significant difference between men and women in the degree of downward pitching among DALL·E 2 images (H3b), and this pattern did not vary by image source. That said, the conditional two-way interaction *Gender* × *Female Dominated* was statistically significant, $b = -3.60$, $p = .023$, suggesting that in DALL·E 2 images, women exhibited a stronger tendency to pitch downward more than men, particularly in *female-dominated* jobs as compared to *male-dominated* jobs. H4b was supported. Regarding RQ2b, no evidence was found to support the three-way interaction.

Discussion

Building upon previous algorithm auditing research documenting the prevalence of gender biases in Google Images (Kay et al., 2015; Lam et al., 2018; Metaxa et al., 2021) and the literature on mass and social media gender biases (Döring et al., 2016; McArthur & Resko, 1975; Parker et al., 2017), we employed a computational approach to empirically examine occupational gender biases in DALL·E 2, an increasingly popular image-generative AI model released by OpenAI. Generative AI models such as DALL·E 2 hold the potential to revolutionize media content production, thereby posing a significant risk of reshaping the media landscape in biased ways upon under-scrutinized adoption. After comparing DALL·E 2 with Google Images and the 2021 U.S. census labor statistics across 153 occupations, we found that DALL·E 2 risks amplifying representational gender bias and two cases of presentational gender biases (i.e., smiling and face pitching). Based on these findings, we encourage future research to study both short-term and long-term impacts of exposure to AI-generated gender stereotyping, particularly a wider range of presentational biases that are understudied in previous computational research on media gender biases. Given the lack of transparency in the model training and development process of DALL·E 2, these findings highlight the need for continuous monitoring of gender biases in generative AI technologies. This requires collaborative efforts from researchers, industry professionals, regulators, and the public.

First, DALL·E 2 systematically underrepresented women in *male-dominated* jobs, while it overrepresented women in images portraying *female-dominated* occupations. This is consistent with prior research that has documented similar representational gender bias in Google Images (Kay et al., 2015; Lam et al., 2018), with one important deviation: In our study, DALL·E 2 exhibited more severe representational bias than Google Image. Ramesh et al. (2021) reported that DALL·E 2 used "text-image pairs from the internet" (p. 4) for training but did not provide details on how the training dataset was constructed. This lack of transparency persists in the recent publication of the "GPT-4 Technical Report" by OpenAI (2023), in which OpenAI attributes biases present in DALL·E 2 generated images to existing biases within their current training data.[5] However, our results revealed that Google Images, regarded as the most commonly used online image search engine, displayed less representational bias than DALL·E 2, with fewer instances of underrepresentation and overrepresentation of women across most occupations. Given the striking differences in representational bias between DALL·E 2 and Google Images documented in our study, DALL·E 2's biases might not be attributed solely to online image data sources like Google Images. The origin of representational gender bias seems to go beyond using training data of "images from the internet," at least beyond those sourced from Google Images.

Given the "black box" nature of generative AI models such as DALL·E 2, pinpointing and mitigating biases is challenging. Therefore, the generative AI industry should seek to establish a collaboration protocol with the academic community for data

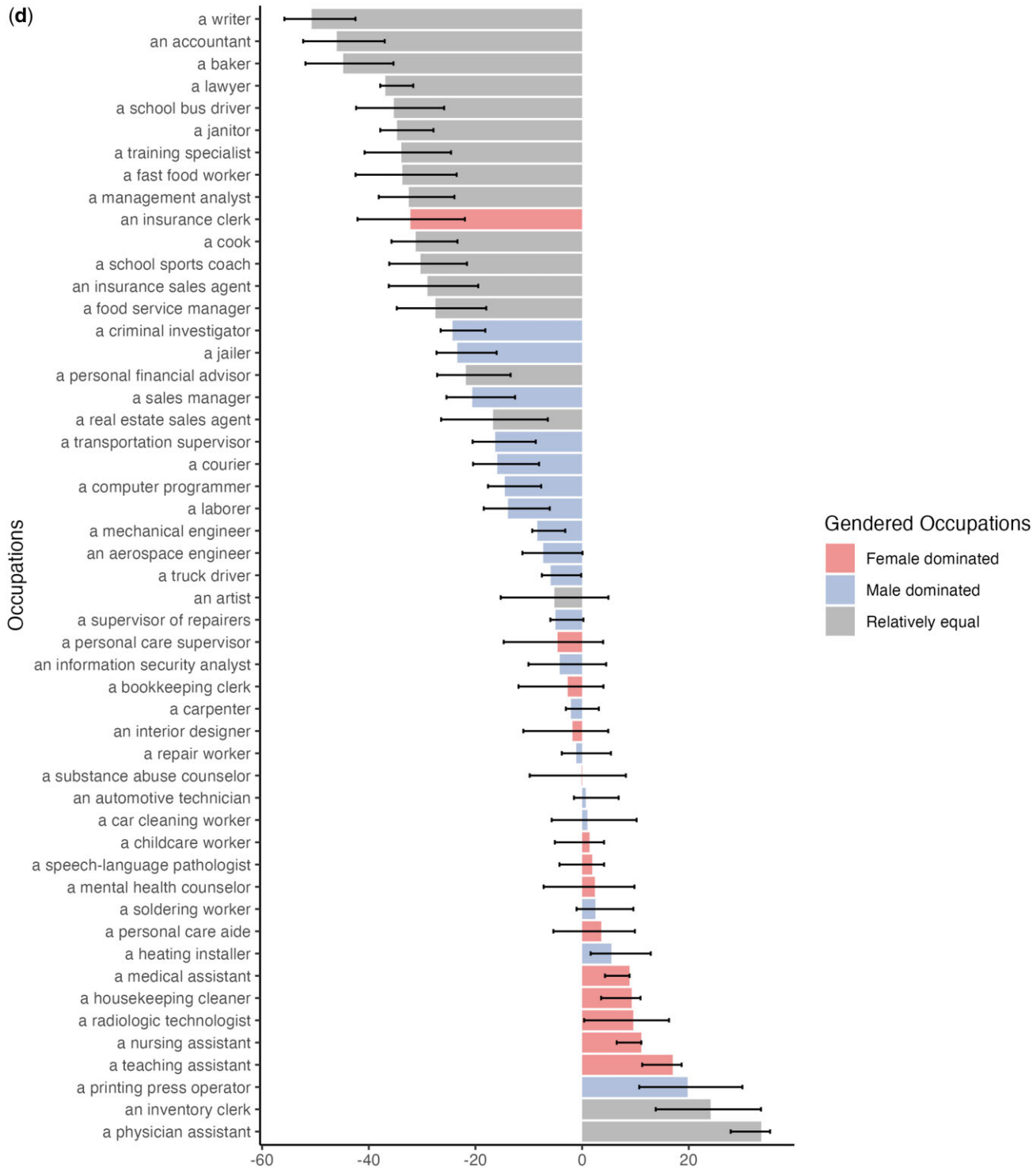

Figure 2 (D). (Continued).

sharing, model performance auditing, and algorithm debiasing in ways that are safe and ethically grounded. Only by enhancing the transparency of AI products can the public, AI professionals, and researchers from different fields engage in collective decision-making processes. This will help prevent the monopoly of powerful stakeholders on AI technologies and reverse the trend of exacerbating existing gender biases.

Second, we went beyond representational bias, the typical focus of prior research on algorithmic gender biases (e.g., Kay et al., 2015; Lam et al., 2018), to examine presentational biases, including smiling and face-pitching. These two visual features are conceptually related to the warmth and competence dimensions highlighted in the stereotype content model (Fiske et al., 2002), and have been studied in existing research

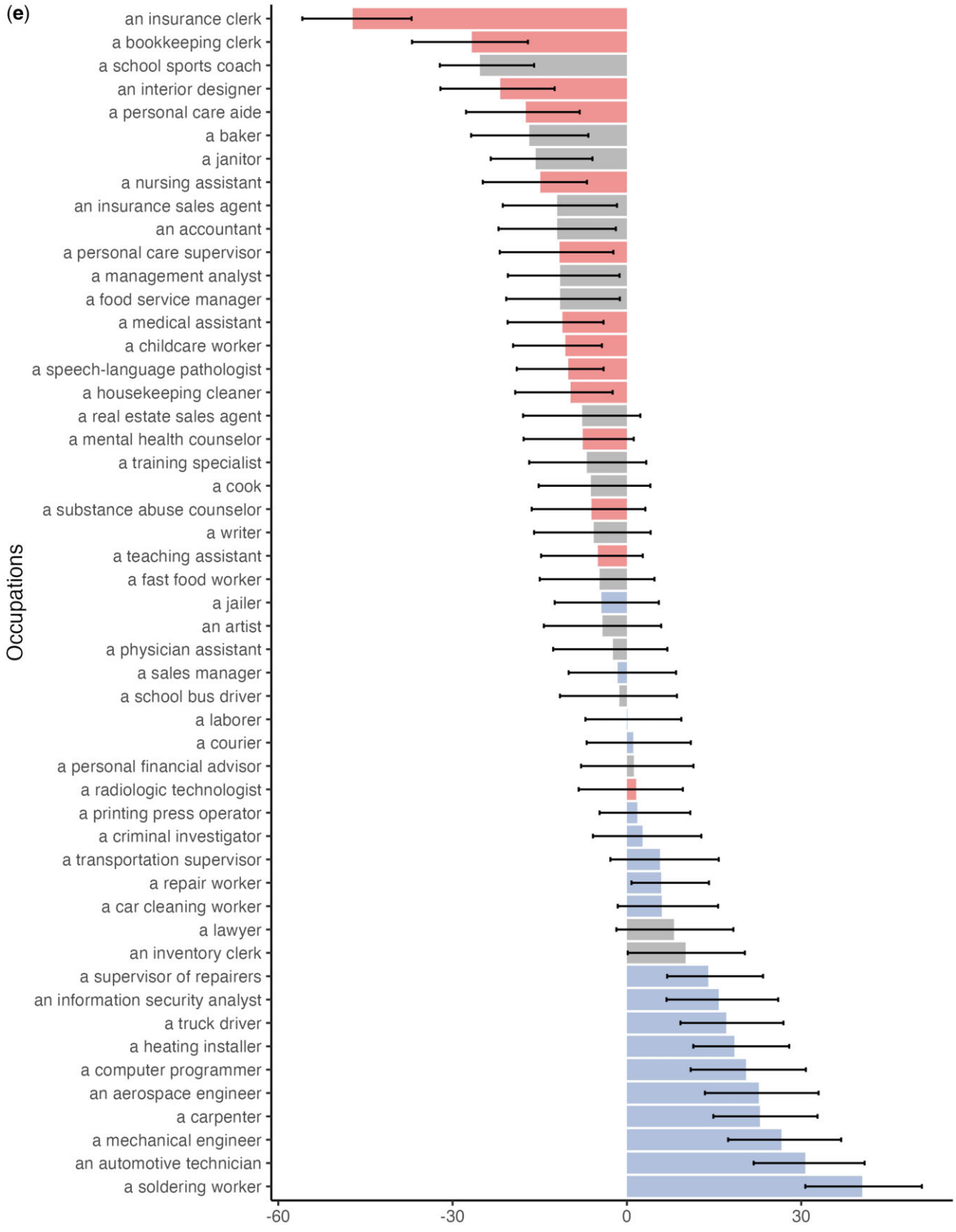

Figure 2 (E). (Continued).

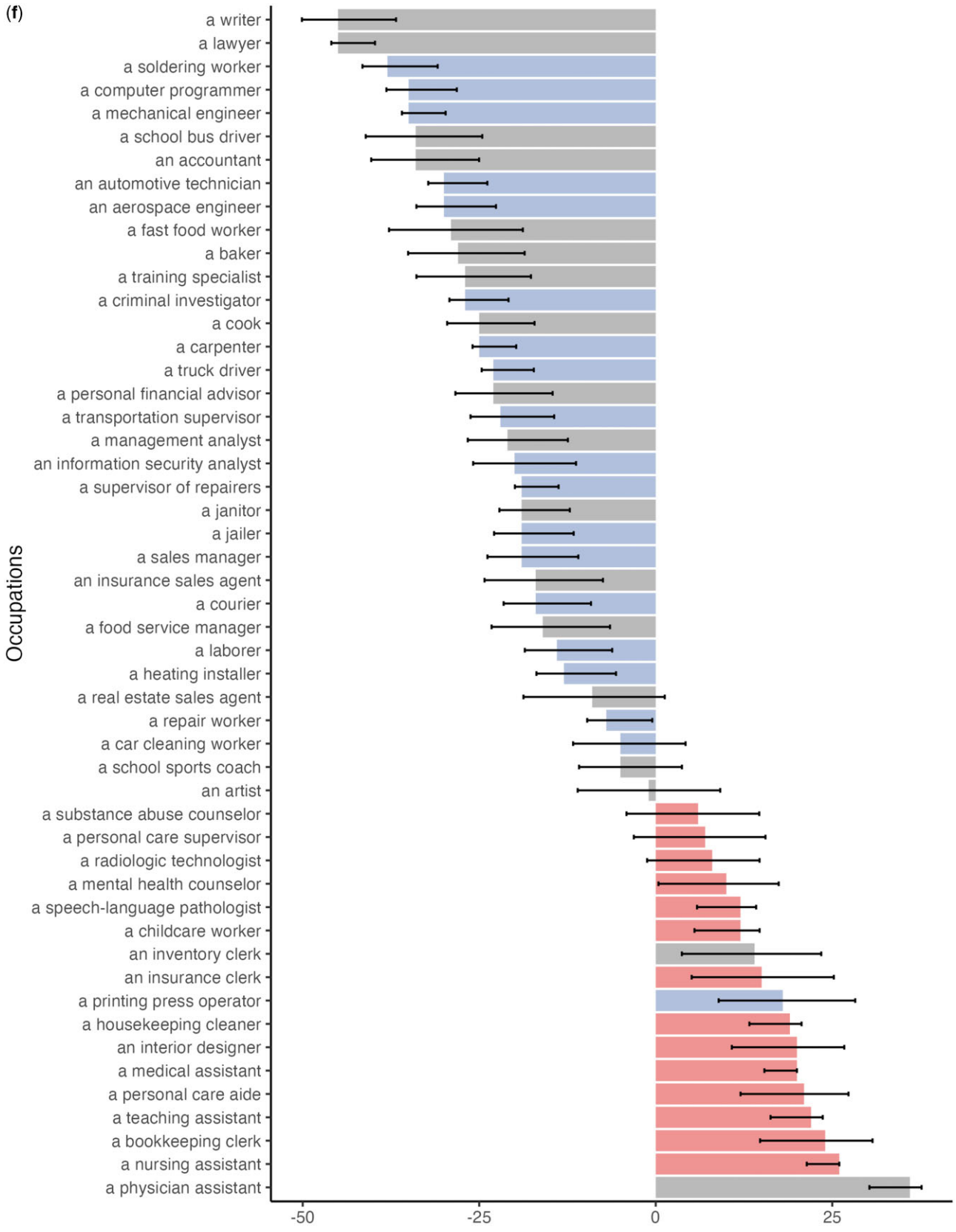

Figure 2 (F). (Continued).

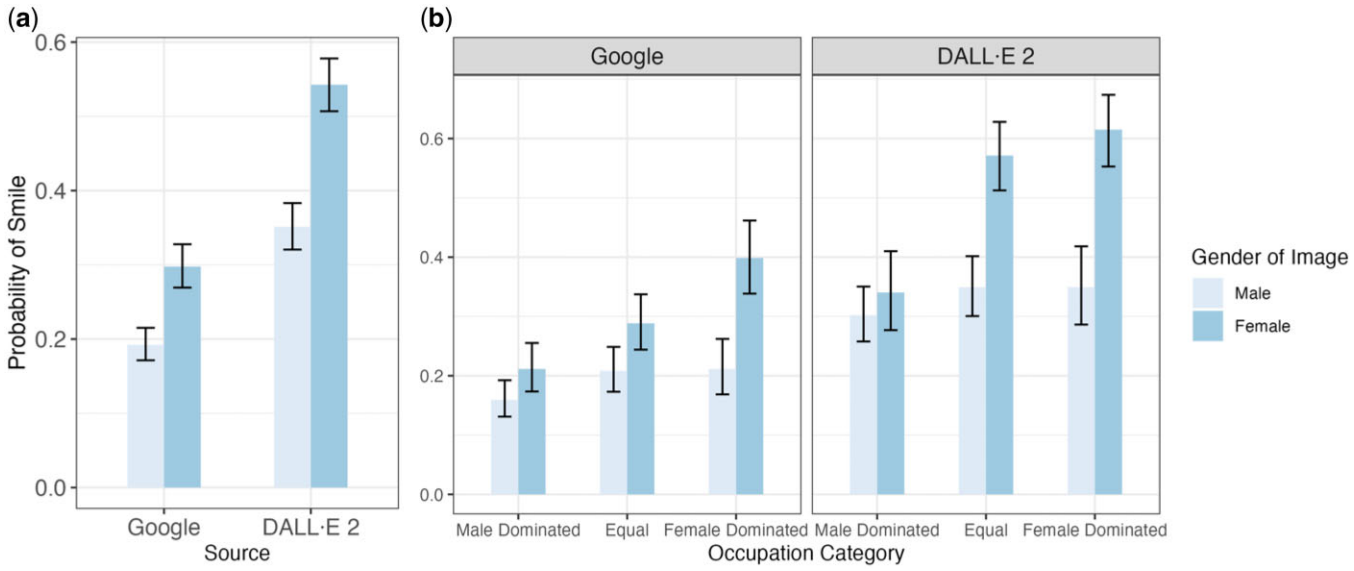

Figure 3. Probability of smile. (A) Source and (B) Occupation category.

Note. Panel (A) presents results from Table 2, Model 1; panel (B) presents results from Table 2, Model 2. Error bars are 95% confidence intervals.

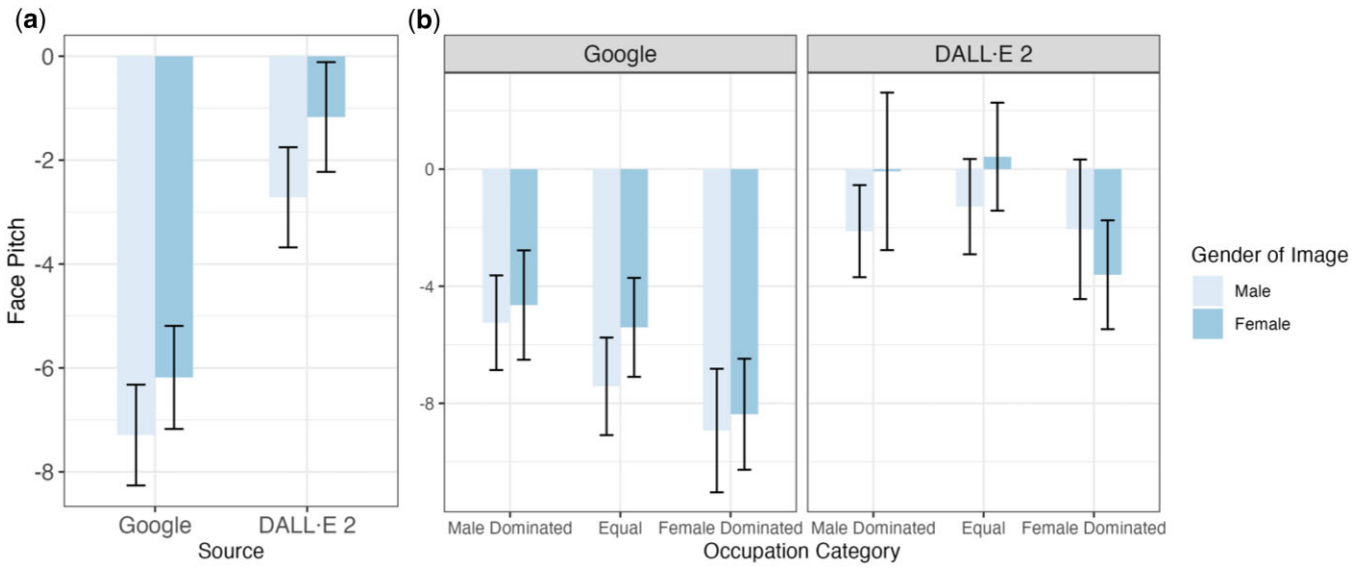

Figure 4. Face pitch. (A) Source and (B) Occupation category.

Note. Panel (A) presents results from Table 3, Model 1; panel (B) presents results from Table 3, Model 2. Error bars are 95% confidence intervals.

on mass and social media gender biases (Collins, 2011; Grau & Zotos, 2016; Peng, 2018; Renner & Masch, 2019; Rodgers et al., 2007). We found that across occupation categories, DALL·E 2 produced more images portraying smiling women, and such gender disparities in facial emotions were more severe than in Google Images. Furthermore, DALL·E 2 was more likely to present smiling women and women whose heads pitch down in *female-dominated* (versus *male-dominated*) occupations, demonstrating the risk of amplifying presentational gender biases. Consistent with priming effects (Jo & Berkowitz 1994; Domke et al., 1998), Peng (2018) finds that different portrayals of facial expressions of Hillary Clinton versus Donald Trump can influence impression formation. Similar priming effects have also been documented in Google Images search results, linking representational biases to the perceptions of occupational gender disparities (Kay et al., 2015; Metaxa et al., 2021). Therefore, although our study is descriptive in nature and cannot speak to the consequences of cumulative exposure to such presentational biases, we have reasons to believe that once widely circulated, DALL·E 2-generated images are likely to reinforce the stereotype of emotional women versus rational men (Renner & Masch, 2019) and further entrench the relationship between femininity and submission (Rose et al., 2012). We encourage future research to examine the "effects" of exposure to AI-generated gendered occupational images or active participation in human-AI co-production of bias-laden media content.

Lastly, given the potential for AI models to perpetuate various biases, recent works have raised questions about the ethical responsibilities of online information providers and the

ways to combat the reproduction of inequalities (Hofeditz et al., 2022; Quadflieg et al., 2022). Proposed solutions include legislative approaches, administrative regulations, nondiscrimination by code or design, and ethics guidelines. Quadflieg et al. (2022) also called for individuals to act upon disobedience to resist the negative effects of AI power. However, few studies have integrated feminist epistemology into their analysis of AI gender bias. In her work examining past and current practices concerning feminist AI, Toupin (2023) argues that feminist knowledge production within AI has been undervalued due to the prevailing rationalist paradigm of "male as norm." Furthermore, feminist perspectives were excluded from AI history, resulting in a masculinist and rationalist historical account. Therefore, Toupin calls for rethinking AI with feminist epistemology and offering alternative narratives to challenge the *status quo*.

Additionally, Wellner and Rothman (2020) review four strategies for eliminating gender bias in AI: ignoring gender references, revealing algorithmic decision-making considerations, designing non-biased algorithms, and involving humans in the process. They advocate for increasing awareness of gender bias and eliminating it based on the feminist understanding that visibility matters. In light of this thinking, it is of critical importance for developers and other stakeholders in the generative AI industry to take responsibility and incorporate feminist epistemology into their daily practices to reduce gender biases.

Limitations

This study has several limitations worth noting. First, we focused on the most prominent face in the multi-face images and followed the default classification threshold (50%) for face and gender detection, which may have introduced measurement errors into our analyses. That said, our extensive crowdsourcing results showed acceptable validity of machine annotations from the AWS *Rekognition* API, and our robustness analyses excluding multi-face images largely replicated the main findings. Therefore, we believe measurement errors are unlikely to confound our key conclusions, although future research is encouraged to employ crowdsourcing to extract visual features when resources permit. Second, we treated gender as a dichotomous variable due to the limitations of the AWS *Rekognition* algorithm and the CPS census data. Future research should seek to expand gender categories to improve inclusivity. Third, direct replication of our study may be challenging given the stochastic nature of generative AI technologies and constant updates to the "black box" of the underlying generative model by OpenAI. To support open science practices, we have uploaded all the image stimuli to OSF, along with codes and the questionnaire for crowdsourcing. That said, whether our results can be replicated with future iterations of DALLE·2 or generalizable to other generative AI technologies (e.g., Midjourney) remains to be seen. We emphasize the need for regular auditing of both representational and presentational gender biases in AI-generated media content, perhaps through broader collaboration across multiple research teams. Fourth, regarding the context of this study, we focused on the CPS census in the United States as the benchmark for comparison. We encourage other researchers to conduct similar AI-auditing research in a more global context to assess generalizability. Fifth, the large size of our image stimuli pool makes it infeasible to rely on crowdsourcing for race and ethnicity classification. Given the lack of valid automatic tools for measurement, we did not pursue intersectionality analyses. Examining how gender biases may compound other dimensions of bias in AI-generated content is an important direction for future research. Finally, although our study revealed gender biases in text-to-image-generative AI from various perspectives, we did not conduct experimental studies to examine how exposure to such gendered images may affect people's perceptions of occupational gender norms and downstream beliefs, attitudinal, and behavioral consequences. We encourage future research to fill this gap and test the effectiveness of potential debiasing strategies such as technical, legal, administrative, and individual resistance approaches (Wellner & Rothman, 2020).

Conclusion

Drawing on the longstanding literature on media gender stereotyping and recent studies documenting occupational gender biases in Google Images search results, this descriptive study reveals that DALL·E 2, a popular image-generative AI model, systematically underrepresents women in male-dominated occupations and overrepresents them in female-dominated jobs. Furthermore, DALL·E 2 images tend to portray more women than men with smiling faces and faces pitching down, particularly in female-dominated (versus male-dominated) occupations, which risks reinforcing traditional gender stereotypes. Our computational algorithm auditing study thus demonstrates the presence of both representational and presentational gender biases in DALL·E 2 images, to a degree more severe than Google Images. These findings emphasize the importance of studying presentational gender biases and the need for continuous monitoring and evaluation of gender biases in generative AI technologies. Future research should expand the scope of gender categories, examine the potential effects of exposure to gendered AI-generated images, and explore strategies to effectively mitigate gender biases in AI models.

Open science framework badges

Open Materials
The components of the research methodology needed to reproduce the reported procedure and analysis are publicly available for this article.

Open Data
Digitally shareable data necessary to reproduce the reported results are publicly available for this article.

Notes

1. See https://www.bls.gov/cps/tables.htm#annual: Labor Force Statistics from the CPS.
2. See https://www.census.gov/topics/population/age-and-sex/about.html and https://www2.census.gov/programs-surveys/cps/techdocs/questionnaires/Demographics.pdf: according to the Bureau of Census, the census uses the concept "sex" rather than "gender" in the questionnaire:
3. See https://serpapi.com: SerpAPI.
4. See https://aws.amazon.com/rekognition: Amazon AWS *Rekognition*.
5. See https://openai.com/research/dall-e-2-pre-training-mitigations.

Supplementary material

Supplementary material is available at *Journal of Computer-Mediated Communication* online.

Data availability

Replication data and code can be accessed at: https://doi.org/10.17605/OSF.IO/3B8EV

Funding

Support for this research was provided by the University of Wisconsin–Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation, awarded to Sijia Yang (#MSN275569 and #MSN275781).

Conflicts of interest: None declared.

Acknowledgments

The authors express gratitude to Dr. Danaë Metaxa for their constructional suggestions, and to Lily Anne Hankley, Meisi Li, and Xinlin Jiang for their research assistance.

References

Appel, M., & Weber, S. (2021). Do mass mediated stereotypes harm members of negatively stereotyped groups? A meta-analytical review on media-generated stereotype threat and stereotype lift. *Communication Research*, *48*(2), 151–179. https://doi.org/10.1177/0093650217715543

Bartoletti, I. (2020). *An artificial revolution: On power, politics and AI*. The Indigo Press.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*, 4349–4357. https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Brandao, M. (2019). *Age and gender bias in pedestrian detection algorithms*. arXiv. https://doi.org/10.48550/arXiv.1906.10490

Brescoll, V. L. (2016). Leading with their hearts? How gender stereotypes of emotion lead to biased evaluations of female leaders. *The Leadership Quarterly*, *27*(3), 415–428. https://doi.org/10.1016/j.leaqua.2016.02.005

Buolamwini, J. A. (2017). *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers* [Doctoral dissertation]. Massachusetts Institute of Technology.

Cappelli, P., Tambe, P., & Yakubovich, V. (2018). Artificial intelligence in human resources management: Challenges and a path forward. *SSRN Electronic Journal*, *61*(4), 15–42. https://doi.org/10.2139/ssrn.3263878

Collins, R. L. (2011). Content analysis of gender roles in media: Where are we now and where should we go? *Sex Roles*, *64*(3), 290–298. https://doi.org/10.1007/s11199-010-9929-5

Coltrane, S., & Adams, M. (1997). Work–family imagery and gender stereotypes: Television and the reproduction of difference. *Journal of Vocational Behavior*, *50*(2), 323–347. https://doi.org/10.1006/jvbe.1996.1575

Davies, P., Quinn, D., & Gerhardstein Nader, R. (2002). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin*, *28*(12), 1615–1628. https://doi.org/10.1177/014616702237644

Domke, D., Shah, D. V., & Wackman, D. B. (1998). Media priming effects: Accessibility, association, and activation. *International Journal of Public Opinion Research*, *10*(1), 51–74. https://doi.org/10.1093/ijpor/10.1.51

Döring, N., Reif, A., & Poeschl, S. (2016). How gender-stereotypical are selfies? A content analysis and comparison with magazine adverts. *Computers in Human Behavior*, *55(B)*, 955–962. https://doi.org/10.1016/j.chb.2015.10.001

Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of Communication*, *57*(1), 163–173. https://doi.org/10.1111/j.1460-2466.2006.00336.x

Fiske, S. T., Cuddy, A. J. C., Glick, P., & 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. https://doi.org/10.1037/0022-3514.82.6.878

Gibbs, S. (2015, July 8). Women less likely to be shown ads for high-paid jobs on google, study shows google the guardian. *The Guardian*. https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

Good, J. J., Woodzicka, J. A., & Wingfield, L. C. (2010). The effects of gender stereotypic and counter-stereotypic textbook images on science performance. *The Journal of Social Psychology*, *150*(2), 132–147. https://doi.org/10.1080/00224540903366552

Grabe, M. E., & Bucy, E. P. (2009). *Image bite politics: News and the visual framing of elections* (1st ed.). Oxford University Press.

Grau, S., & Zotos, Y. (2016). Gender stereotypes in advertising: A review of current research. *International Journal of Advertising*, *35*(5), 761–770. https://doi.org/10.1080/02650487.2016.1203556

Hamilton, M. C., Anderson, D., Broaddus, M., & Young, K. (2006). Gender stereotyping and under-representation of female characters in 200 popular children's picture books: A twenty-first century update. *Sex Roles*, *55*(11), 757–765. https://doi.org/10.1007/s11199-006-9128-6

Hofeditz, L., Mirbabaie, M., Luther, A., Mauth, R., & Rentemeister, I. (2022). Ethics guidelines for using AI-based algorithms in recruiting: Learnings from a systematic literature review. *Proceedings of the 55th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2022.018

Jerald, M. C., Ward, L. M., Moss, L., Thomas, K., & Fletcher, K. D. (2017). Subordinates, sex objects, or sapphires? Investigating contributions of media use to Black students' femininity ideologies and stereotypes about Black women. *Journal of Black Psychology*, *43*(6), 608–635. https://doi.org/10.1177/0095798416665967

Jo, E., & Berkowitz, L. (1994). A priming effect analysis on media influences: An update. In J. Bryant & D. Zillman (Eds.), *Media effects: Advances in theory and research* (pp. 43–60). Erlbaum.

Johnston, I. (2017, April 14). AI robots learning racism, sexism and other prejudices from humans, study finds. *The Independent*. https://www.independent.co.uk/tech/ai-robots-artificial-intelligence-racism-sexism-prejudice-bias-language-learn-from-humans-a7683161.html

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3819–3828). https://doi.org/10.1145/2702123.2702520

Kerkhoven, A. H., Russo, P., Land-Zandstra, A. M., Saxena, A., & Rodenburg, F. J. (2016). Gender stereotypes in science education resources: A visual content analysis. *PLoS One*, *11*(11), e0165037. https://doi.org/10.1371/journal.pone.0165037

Kirk, H. R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A., & Asano, Y. (2021). Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in Neural Information Processing*

Systems, *34*, 2611–2624. https://proceedings.neurips.cc/paper/2021/hash/1531beb762df4029513ebf9295e0d34f-Abstract.html

Knoll, S., Eisend, M., & Steinhagen, J. (2011). Gender roles in advertising: A comparison of gender stereotyping on public and private TV channels in Germany. *International Journal of Advertising*, *30*(5), 867–888. https://doi.org/10.2501/IJA-30-5-867-888

LaFrance, A. (2016, March 30). Why do so many digital assistants have feminine names? *The Atlantic*. https://www.theatlantic.com/technology/archive/2016/03/why-do-so-many-digital-assistants-have-feminine-names/475884/

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, *65*(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

Lam, O., Broderick, B., Wojcik, S., & Hughes, A. (2018, December 17). *Gender and jobs in online image searches*. Pew Research Center. https://www.pewresearch.org/social-trends/2018/12/17/gender-and-jobs-in-online-image-searches/

Lawton, G. (2023, March). What is generative AI? Everything you need to know. *TechTarget*. Retrieved March 18, 2023. https://www.techtarget.com/searchenterpriseai/definition/generative-AI

Lucy, L., & Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. *Proceedings of the Third Workshop on Narrative Understanding*, 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5

McArthur, L. Z., & Resko, B. G. (1975). The portrayal of men and women in American television commercials. *The Journal of Social Psychology*, *97*(2), 209–220. https://doi.org/10.1080/00224545.1975.9923340

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Metaxa, D., Gan, M. A., Goh, S., Hancock, J., & Landay, J. A. (2021). An image of society: Gender and racial representation and impact in image search results for occupations. *Proceedings of the ACM on Human-Computer Interaction*, *5(CSCW1)*, 1–23. https://doi.org/10.1145/3449100

Min, X., Zhai, G., Gu, K., Liu, J., Wang, S., Zhang, X., & Yang, X. (2017). Visual attention analysis and prediction on human faces. *Information Sciences*, *420*, 417–430. https://doi.org/10.1016/j.ins.2017.08.040

Monea, A. (2022). *The digital closet: How the internet became straight*. https://doi.org/10.7551/mitpress/12551.001.0001

Nagpal, S., Singh, M., Singh, R., & Vatsa, M. (2019). *Deep learning for face recognition: Pride or prejudiced?* arXiv preprint arXiv: 1904.01219. https://doi.org/10.48550/arXiv.1904.01219

Noelle-Neumann, E. (1993). *The spiral of silence: Public opinion - Our social skin* (2nd ed.). The University of Chicago Press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, *2*, 13. https://doi.org/10.3389/fdata.2019.00013

OpenAI. (2022, June 28). *DALL·E 2 pre-training mitigations*. OpenAI. https://openai.com/research/dall-e-2-pre-training-mitigations

OpenAI. (2023). *GPT-4 technical report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Parker, R., Larkin, T., & Cockburn, J. (2017). A visual analysis of gender bias in contemporary anatomy textbooks. *Social Science & Medicine*, *180*, 106–113. https://doi.org/10.1016/j.socscimed.2017.03.032

Peng, Y. (2018). Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, *68*(5), 920–941. https://doi.org/10.1093/joc/jqy041

Plakoyiannaki, E., Mathioudaki, K., Dimitratos, P., & Zotos, Y. (2008). Images of women in online advertisements of global products: Does sexism exist? *Journal of Business Ethics*, *83*(1), 101–112. https://doi.org/10.1007/s10551-007-9651-6

Plant, E. A., Hyde, J. S., Keltner, D., & Devine, P. G. (2000). The gender stereotyping of emotions. *Psychology of Women Quarterly*, *24*(1), 81–92. https://doi.org/10.1111/j.1471-6402.2000.tb01024.x

Quadflieg, S., Neuburg, K., & Nestler, S. (Eds.). (2022). *(Dis)Obedience in digital societies: Perspectives on the power of algorithms and data*. transcript Verlag. https://doi.org/10.1515/9783839457634

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021, July). Zero-shot text-to-image generation. *International Conference on Machine Learning* (pp. 8821–8831). PMLR. https://proceedings.mlr.press/v139/ramesh21a.html

Renner, A. M., & Masch, L. (2019). Emotional woman – rational man? Gender stereotypical emotional expressivity of German politicians in news broadcasts. *Communications*, *44*(1), 81–103. https://doi.org/10.1515/commun-2017-0048

Ridgeway, C. L. (2001). Gender, status, and leadership. *Journal of Social Issues*, *57*(4), 637–655. https://doi.org/10.1111/0022-4537.00233

Robertson, R. E., Lazer, D., & Wilson, C. (2018, April). Auditing the personalization and composition of politically-related search engine results pages. *Proceedings of the 2018 World Wide Web Conference* (pp. 955–965). https://doi.org/10.1145/3178876.3186143

Rodgers, S., Kenix, L. J., & Thorson, E. (2007). Stereotypical portrayals of emotionality in news photos. *Mass Communication and Society*, *10*(1), 119–138. https://doi.org/10.1080/15205430709337007

Rose, J., Mackey-Kallis, S., Shyles, L., Barry, K., Biagini, D., Hart, C., & Jack, L. (2012). Face it: The impact of gender on social media images. *Communication Quarterly*, *60*(5), 588–607. https://doi.org/10.1080/01463373.2012.725005

Seiter, E. (1986). Stereotypes and the media: A re-evaluation. *Journal of Communication*, *36*(2), 14–26. https://doi.org/10.1111/j.1460-2466.1986.tb01420.x

Sherman, S.J., Mackie, D.M. & Driscoll, D.M. (1990). Priming and the differential use of dimensions in evaluation. *Personality and Social Psychology Bulletin*, *16*(3), 405–418. https://doi.org/10.1177/0146167290163001

Shor, E., van de Rijt, A., Miltsov, A., Kulkarni, V., & Skiena, S. (2015). A Paper ceiling: Explaining the persistent underrepresentation of women in printed news. *American Sociological Review*, *80*(5), 960–984. https://doi.org/10.1177/0003122415596999

Singh, V., Chayko, M., Inamdar, R., & Floegel, D. (2020). Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms. *Journal of the Association for Information Science and Technology*, *71*(11), 1281–1294. https://doi.org/10.1002/ASI.24335

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and access in algorithms, mechanisms, and optimization* (pp. 1–9). https://doi.org/10.1145/3465416.3483305

Toupin, S. (2023). Shaping feminist artificial intelligence. *New Media & Society*, *0*(0). https://doi.org/10.1177/14614448221150776

Vlasceanu, M., & Amodio, D. M. (2022). Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences*, *119*(29), e2204529119. https://doi.org/10.1073/pnas.2204529119

Wellner, G., & Rothman, T. (2020). Feminist AI: Can we expect our AI systems to become feminist? *Philosophy & Technology*, *33*(2), 191–205. https://doi.org/10.1007/s13347-019-00352-z

Wiggers, K. (2022, November 3). *Now anyone can build apps that use DALL-E 2 to generate images*. TechCrunch. https://techcrunch.com/2022/11/03/now-anyone-can-build-apps-that-use-dall-e-2-to-generate-images/

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. *Advances in experimental social psychology* (Vol. *46*, pp. 55–123). Elsevier. https://doi.org/10.1016/B978-0-12-394281-4.00002-7