

Text-to-image models globally default to male, enforce gender stereotypes, and penalize female and nonbinary representations

Genevieve Smith

genevieve.smith@berkeley.edu

University of California, Berkeley <https://orcid.org/0000-0002-0615-6378>

Leander Girkbach

Munich Center for Machine Learning (MCML); Technical University of Munich

Vongani Maluleke

University of California, Berkeley

Sidhika Balachandar

University of California, Berkeley

Zeynep Akata

Munich Center for Machine Learning (MCML); Technical University of Munich

Stephan Alaniz

LTCI, Télécom Paris, Institut Polytechnique de Paris

Article

Keywords:

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11435572/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

There is **NO** Competing Interest.

This work received ethical approval from UC Berkeley's Office for Protection of Human Subjects (OPHS), study's CPHS454 protocol ID (2025-09-18928).

Text-to-image models globally default to male, enforce gender stereotypes, and penalize female and nonbinary representations

Anonymous Author(s)

Abstract

As text-to-image (T2I) systems generate synthetic imagery at scale, their outputs systematically embed and amplify structural gender biases. We present the largest, most globally representative study of gender bias in T2I generation, analyzing nearly 1 million images across 22 countries and four leading open-weight models, alongside 3,533 human annotations. We identify three interconnected gender logics embedded in generative models. First, androcentrism: gender-neutral prompts yield male representation in 68.6% of outputs—a default beyond real-world composition that increases in high-power roles and countries with higher gender inequality; while images of women incur quality penalties. Second, bidirectional stereotyping: men are portrayed as fit, muscular, and stern; and women as young, thin, attractive, and smiling. Third, nonbinary marginalization: models show reduced comprehension when prompted to depict nonbinary people. Annotators report discomfort, exclusion, and insecurity viewing homogeneous depictions of their identity, while counter-stereotypical images can feel affirming.

1 Introduction

Text-to-image (T2I) generative models are a pervasive source of visual content, increasingly deployed across advertising, media, education, and everyday creative and professional use, with an estimated 86% of creators using AI for asset creation [1]. Industry estimates indicate over 15 billion images were generated with T2I models between 2022 and 2023 alone. Approximately 80% of these images were created using the open-weight Stable Diffusion ecosystem [2]. The growing use of T2I models gives the visual defaults they encode immense power to shape how people perceive gender, occupation, and identity at scale. Exposure to gendered media representations is well established to strengthen stereotype endorsement and shape career-related ambitions and self-perception [3], while exposure to biased image search results shifts real-world

beliefs [4]. Understanding whether and how T2I systems reproduce societal gender biases is therefore a pressing empirical and ethical concern.

Three questions remain central, yet no study has addressed them comprehensively and at scale. Who do these models depict by default when gender is left unspecified? When a specific gender identity is explicitly requested, how faithfully, realistically, and stereotypically is that identity rendered? How do the people who share these identities experience seeing themselves represented in this way?

On the first question, prior work demonstrates that T2I outputs default to male representation across occupational and neutral persona prompts in narrower samples [5–8], while also amplifying gender disparities beyond real-world labor statistics [9–11]. Whether such male defaults and occupational skews hold at global scale, across leading models, and in relation to structural gender inequality remains untested. On the second question, prior work reveals T2I models depict more women with smiles and downward-pitching heads [10], while web imagery and language models depict women as systematically younger than men, an intersectional bias tied to occupational status [4]. Related work has further documented homogenized racial and national depictions in T2I outputs [12, 13]. None of this work assesses whether image quality or realism varies across prompted identities, nor conducts stereotype measurement through identity-matched human evaluation grounded in established social-psychological theory. On the third question, how depicted groups experience these representations, work has been limited to a handful of qualitative interviews [14], leaving open how people who share a depicted identity perceive and are affected by it. Nonbinary representation, across all three questions, remains largely unexamined.

This paper addresses these gaps through two studies. In the first, we generated nearly 1 million images with gender-neutral prompts across 105 occupations and activities. Occupations were balanced across male- and female-dominated roles and across low, medium, and high occupational power (the authority, status, and decision-making scope typically attached to a role). Activities were adapted from a classification of masculine-, feminine-, and neutral-coded activities [15], slightly skewed towards female-coded activities. We focus on open-weight models because they represent a disproportionate share of image generation [2], are surging in adoption globally [16, 17], and our broader research program includes bias mitigation via direct model fine-tuning (see Methods for details on occupation, activity, and model selection). From this, we analyzed gender representation in the outputs using automated classification (see ‘Automated classification study’ in Methods). In the second study, we generated gender- and nationality-specific images for a balanced subset of 10 occupations and 10 activities drawn from the same pool (preserving the same gender-balance and power-level balance), across seven nationalities representing each world region for two of the four T2I models considered. Human annotators then evaluated randomly sampled images matching their own gender and nationality, yielding 3,533 human annotations across a structured, multi-metric protocol (see ‘Human annotation study’ in Methods) allowing us to assess not only who is represented, but how well they are rendered, whether their depiction reflects established gender-related physical and trait-based stereotypes, and how their representation is experienced by the people it depicts. Across both studies, our findings reveal that T2I outputs embed three gender

logics: androcentrism [18, 19] reflected in male defaults (Result 1) and quality penalties for female depictions (Result 2); bidirectional stereotyping for men and women [20] (Result 3); and nonbinary marginalization manifested in lower model comprehension for nonbinary prompts (Result 4), consistent with accounts of algorithmic gender illegibility [21, 22].

2 Results

2.1 Male defaults dominate gender-neutral prompts

Across gender-neutral prompts specifying nationality (e.g., “an Indian person working as a translator”), T2I models default to depicting men more often than women. Averaged across the 22 nationalities and four models, the mean male ratio was 68.6%, despite occupation and activity prompts deliberately balanced such that a model reflecting real-world statistics would produce outputs close to an even male/female split overall.

This default varied substantially by nationality (see Fig. 1a). It was most pronounced for Uzbekistani (86.6%) and Indian (82.5%) prompts, and weakest for French Polynesian (53.2%) and Chinese prompts (44.1%). Chinese prompts were the only nationality that skewed towards female outputs. A manual review confirmed this reversal reflects genuine model behavior rather than classification error (see Methods). The male-default pattern was present across all four models tested, with varying magnitude. Flux-Schnell showed the strongest overall male ratio (74.2% average across nationalities), followed by SD3.5-Large (70.5%), Lumina2 (67.6%), and Flux (61.9%). The relative ranking of nationalities was broadly consistent across models (see Fig. 1b).

The default varied by structural context. Nationalities of low- and middle-income countries (LMICs) showed higher male ratio (77.7%) relative to high-income countries (67.2%; $t = 83.22$, $P < 0.001$); income groups follow World Bank classifications, excluding Guam and French Polynesia, which are territories without independent classifications. A parallel pattern held for country-level gender inequality, measured with the OECD’s Social Institutions and Gender Index (SIGI), which captures gender discrimination embedded in laws, social norms, and practices. Grouping nationalities into SIGI’s four standard categories (excluding Uzbekistan, Guam, and French Polynesia, which lack SIGI scores), male ratio was highest among nationalities in countries with high or very high gender inequality (77.2%) relative to very low inequality contexts (67.6%; $t = -68.16$, $P < 0.001$).

The male-default pattern also varied by occupational power. Regressing male output against real-world male representation (employment statistics from International Labour Organization ISCO-08 unit-group data, latest available across countries globally; see Methods) across the 60 occupations (means averaged across 22 nationalities and the four T2I models) showed that outputs tracked the direction of real-world composition, rising at nearly the same rate as real-world male share increased (slope = 0.986, $R^2 = 0.690$, $P < 0.0001$). However, this relationship held at a consistently higher level, with AI outputs averaging 14.8 percentage points more male than real-world rates (see Fig. 2). This shift scaled with occupational power. High-power occupations showed the largest and most consistent shift (mean +21.9 percentage

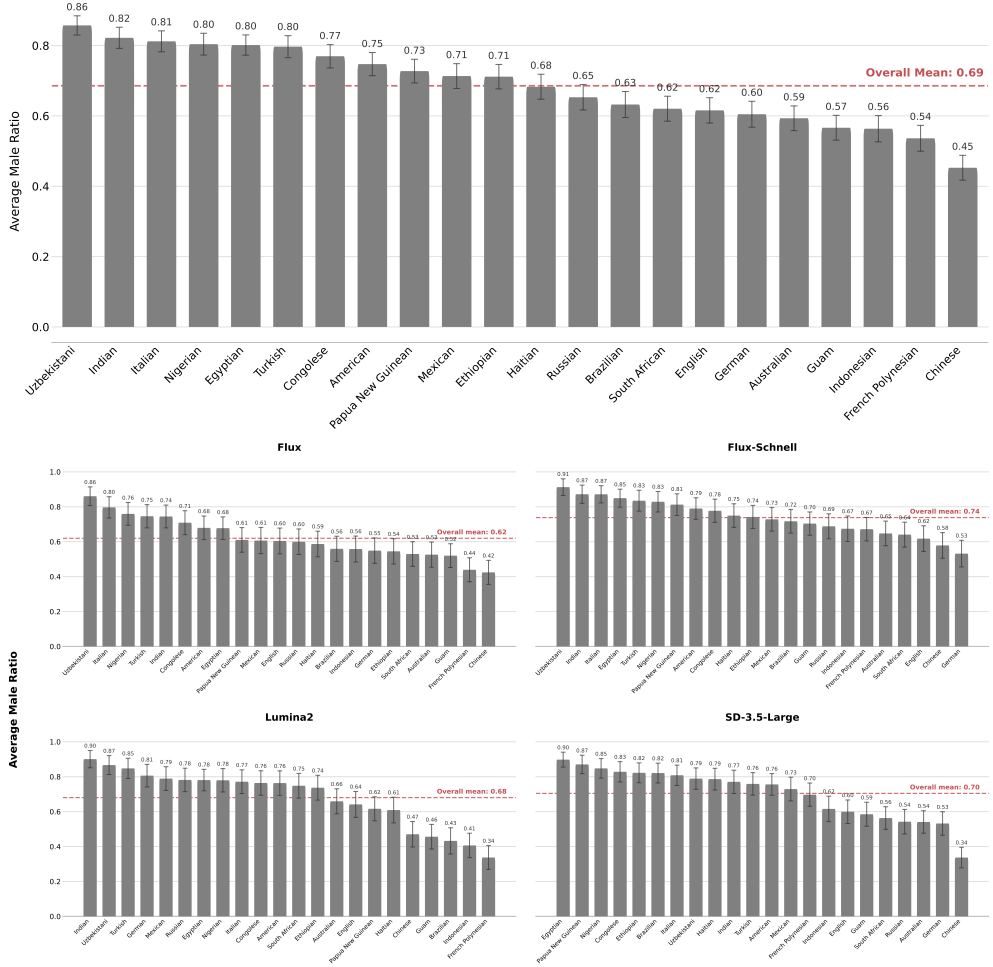


Fig. 1: Male-gender defaults across text-to-image models for nationality-conditioned occupation and activity prompts. **a**, Average proportion of male-labeled images across 22 nationalities, pooled across the four T2I models ($N = 9,240$; $n = 420$ evaluations per nationality). Nationalities are sorted in descending order of male depiction ratio. Outputs are skewed towards men for 21 of 22 nationalities. The horizontal red dashed line shows the overall mean (68.6%), which significantly exceeds gender parity (50%; two-sided one-sample t -test, $t(9239) = 48.14$, $P < 0.001$). **b**, Nationality-specific male depiction ratios broken down across the four models: Flux (62.0%), Flux-Schnell (73.8%), Lumina2 (68.0%), and SD-3.5-Large (70.4%) ($n = 105$ prompts per nationality per model). Horizontal red dashed lines indicate model-wise overall means. Within each panel, nationalities are ordered in descending order of male ratio. In all panels, grey bars denote group means, error bars denote 95% confidence intervals ($\pm 1.96 \times \text{s.e.m.}$, where $\text{s.e.m.} = \text{s.d.}/\sqrt{n}$).

points; 12 of 12 occupations skewed more male than their real-world rate). Medium- and low-power occupations showed smaller positive shifts toward more male representation (mean +14.6 and +11.7 percentage points, respectively). Of the 25 low-power occupations, eight skewed more female than real-world statistics, while 17 skewed more male. These patterns indicate that occupational power informs the extent of T2I model defaults, with high-power roles associated with maleness.

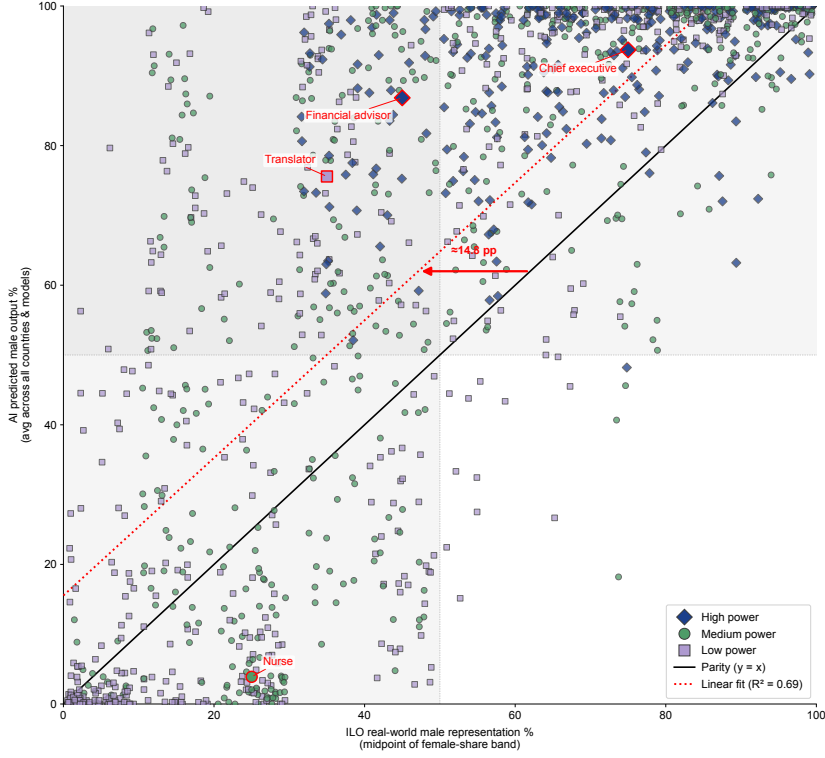


Fig. 2: AI-generated output versus real-world representation, by occupation power status. Each marker is one occupation $\times$ nationality ($n = 1,320$; 60 occupations $\times$ 22 nationalities), giving the percentage of generated images predicted male, pooled across the four T2I models (~ 100 images per model). Marker colour/shape denotes occupational power level: navy diamond, high; green circle, medium; purple square, low. The x-axis is the ILO real-world male share for that occupation (female-share band midpoint), identical across nationalities; x-values are jittered ± 4.5 for visibility. Black line, parity ($y = x$); dotted red line, ordinary-least-squares fit on occupation-level means (one point per occupation, averaged over nationalities and models; $n = 60$): slope = 0.99, $R^2 = 0.69$, $P < 0.0001$; AI output averages 14.8 percentage points more male than the real-world rate. Red-outlined markers highlight four illustrative occupations.

2.2 Female depictions incur quality penalties

Images depicting men received significantly higher quality scores than those depicting women, including increased prompt comprehension (the extent to which the image output reflects an understanding of the prompt, mean male score of 4.0 vs. 3.86 for female), realism of scenario (whether the depicted setting and activity could plausibly occur in the real world, 3.80 vs. 3.65), realism of the person depicted (whether the individual shown resembles a real person of that identity, 3.79 vs. 3.56), and overall perceived quality (3.85 vs. 3.74). Notably, the gap was largest for realism of the person depicted. All items were rated on a 1-5 scale (all $p < 0.01$). Annotator ratings did not differ significantly by rater gender or country on neutral-prompt control images (Supplementary Note 1).

The gap was moderated by national context. In comparing nationalities from high-income countries that have comparatively low structural gender inequality (German, American, Australian) to those from LMICs with high levels of gender inequality (Indian, Egyptian), overall perceived quality scores were comparable, but the latter group received significantly worse comprehension (3.98 vs. 3.56), realism of scenario (3.74 vs. 3.45), and realism of person depicted (3.71 vs. 3.32; all $P < 0.001$), with women from this group scoring worst across every dimension (see Fig. 3).

The gap compounds when occupational gender expectations are violated. Comparing images by gender in prompt (PG) and the occupation’s real-world statistically-dominant gender (SG), men in gender-typical occupations (PG: male, SG: male) scored highest on realism of the person depicted ($M = 3.84$), followed by men in female-dominated occupations ($M = 3.75$, not significantly different from men in male-dominated roles, $p = 0.1565$). Women scored significantly lower than both groups, with women in female-dominated occupations ($M = 3.61$) and most markedly in male-dominated occupations ($M = 3.51$, both $P < 0.05$; see Fig. 4). The largest and most consistent gaps emerge between men in male-dominated occupations and women in male-dominated occupations—most notably for realism of the person depicted (Cohen’s $d = 0.29$, $p < 0.001$) and realism of scenario ($d = 0.20$, $P = 0.003$)—consistent with a counter-stereotypical penalty, whereby generative models render less realistic images of women when they deviate from socially expected gender-occupation norms. This pattern is unlikely to reflect female annotators simply penalizing stereotype violation itself, as images of women scored lower than men in both female-dominated and male-dominated occupation conditions.

2.3 T2I outputs exhibit bidirectional gender stereotyping

T2I outputs reflect bidirectional gender stereotyping, whereby men and women are systematically attributed different traits consistent with traditional masculine and feminine archetypes [20]. Images of women were more likely than images of men to be characterized as attractive (54% vs. 26%), smiling (41% vs. 19%), and thin (40% vs. 20%; all $p < 0.001$, see Fig. 5). Images of men were more likely to be characterized as having a serious or stern expression (36% vs. 23%) and as fit or muscular (32% vs. 15%; both $p < 0.001$; see Fig. 6 for example outputs). Women were also perceived as younger, compared to men (33.4 vs. 40.8 years on average). This pattern of women as

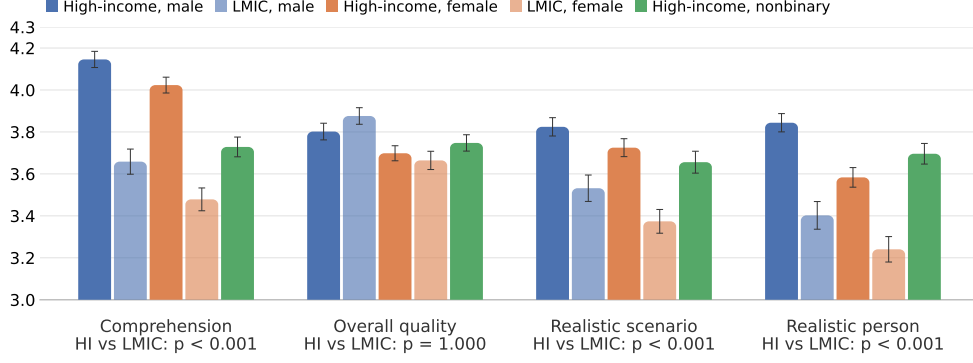


Fig. 3: Depiction quality ratings across gender categories and country income levels. Comparison of annotator quality ratings (1–5 scale) across four dimensions: prompt comprehension, overall perceived quality, realistic scenario, and realistic person. Bars show mean scores across groups: high-income male (dark blue), LMIC male (light blue), high-income female (dark orange), LMIC female (light orange), and high-income nonbinary (green). Nonbinary depictions were collected and evaluated only for high-income nationalities. Error bars indicate ± 1 standard error of the mean (s.e.m.). Sample sizes across all evaluated metrics: high-income male ($n = 703$), LMIC male ($n = 400$), high-income female ($n = 600$), LMIC female ($n = 420$), and high-income nonbinary ($n = 558$), corresponding to total sample sizes of $n = 1,861$ (high-income) and $n = 820$ (LMIC). Statistical differences between high-income and LMIC depictions (across genders) use two-sided Welch’s t -tests with Bonferroni correction for the four tested metrics: comprehension ($t(1353.4) = 8.96$, unadjusted $P < 0.001$, adjusted $P < 0.001$), overall quality ($t(1662.4) = -0.30$, unadjusted $P = 0.764$, adjusted $P = 1.000$), realistic scenario ($t(1426.4) = 5.91$, unadjusted $P < 0.001$, adjusted $P < 0.001$), and realistic person ($t(1370.2) = 7.63$, unadjusted $P < 0.001$, adjusted $P < 0.001$).

169 younger held across each nationality (see Extended Data 7) with the largest age gap
 170 among depicted Americans (30.5 vs. 42.9 years).

171 2.4 Nonbinary prompts yield lower model comprehension

172 Across nationalities with nonbinary annotation data (American, German, Australian),
 173 prompt comprehension scores were significantly worse when nonbinary people were
 174 prompted ($M = 3.73$) than when male ($M = 4.15$) or female ($M = 4.02$) people were
 175 prompted (both $P < 0.001$; see Fig. 3 and Fig. 6). Effect size measured by Cohen’s
 176 d confirm this: $d = 0.4$ for male vs. nonbinary and $d = 0.3$ for female vs. nonbinary,
 177 both amounting to moderate effect sizes. This indicates that generated images are less
 178 likely to reflect the prompted content when a nonbinary identity was specified. This
 179 comprehension penalty did not extend uniformly to other quality dimensions. While
 180 nonbinary prompts scored significantly lower than male prompts on realism of scenario

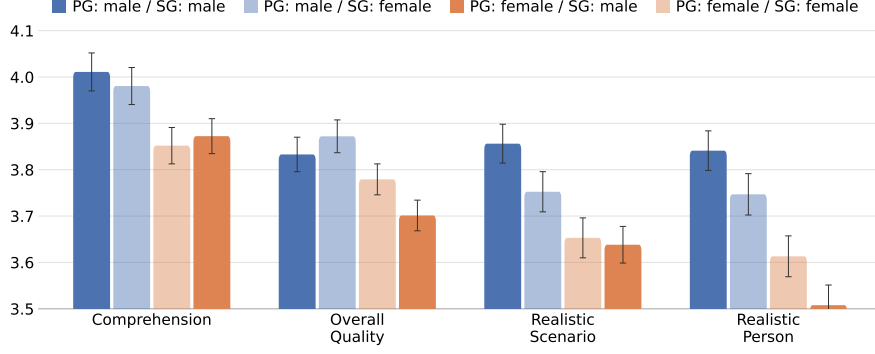


Fig. 4: Quality ratings across prompt gender and stereotypical occupational gender conditions. Comparison of annotator quality ratings (1–5 scale) across four dimensions: prompt comprehension, overall perceived quality, realistic scenario, and realistic person. For each metric, bars represent conditions from left to right: prompt gender (PG) male in a male-dominated occupation (SG: male; dark blue), PG male in a female-dominated occupation (SG: female; light blue), PG female in a female-dominated occupation (SG: female; light orange), and PG female in a male-dominated occupation (SG: male; dark orange). Bar heights denote sample means and error bars indicate ± 1 standard error of the mean (s.e.m.). Sample sizes per condition across evaluated images: PG male / SG male ($n = 767$), PG male / SG female ($n = 756$), PG female / SG female ($n = 735$), and PG female / SG male ($n = 717$). Pairwise differences for the realistic person use two-sided Welch’s t -tests: PG male / SG male vs. PG male / SG female ($t(1443.3) = 1.53$, $P = 0.1274$); PG male / SG male vs. PG female / SG male ($t(1439.7) = 5.47$, $P < 0.001$); PG male / SG male vs. PG female / SG female ($t(1455.5) = 3.71$, $P < 0.001$); and PG female / SG female vs. PG female / SG male ($t(1440.0) = 1.70$, $P = 0.0889$).

and realism of person depicted ($p = 0.0067$ and $P = 0.0123$, respectively), they did not differ significantly from female prompts.

2.5 Depicted groups report both harm and empowerment

Beyond quantitative annotation metrics, open-ended reflections from 173 annotators upon completing image annotations (372 coded excerpts, see Methods) show these patterns are not experienced as abstract statistics. Annotators most frequently described the images as homogenizing, narrow, stereotypical depictions of their identity group. An Indian annotator described feeling, “*a bit strange to see a ‘template’ of your identity. The AI seems to have a very specific idea of what an Indian man looks like (thick beard, dense hair, certain nose shape)... they lack the true diversity of India’s various regions and ethnicities.*” Others, particularly women, pointed to narrow beauty or body standards. One Australian woman described generated images as representing “*the way people, globally, think Australian women look: white but tanned, thin*

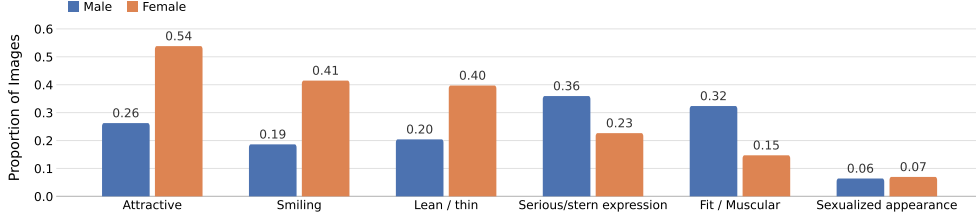


Fig. 5: Gendered divergence in perceived physical and expressive attributes. Proportion of generated images where annotators attribute specific physical, emotional, or stylistic traits across depicted gender categories. Numerical values above each bar state the proportion within that gender category. Total sample sizes are $n = 1,063$ images for male prompts and $n = 1,020$ images for female prompts (excluding images for Australia and Brazil). Differences in attribute attribution rates between male and female depictions use two-sided Pearson’s χ^2 tests (d.f. = 1) with Bonferroni correction for the $m = 6$ tested attributes: attractive ($\chi^2(1) = 165.29$, unadjusted $P < 0.001$, adjusted $P < 0.001$), smiling ($\chi^2(1) = 129.82$, unadjusted $P < 0.001$, adjusted $P < 0.001$), lean / thin ($\chi^2(1) = 92.50$, unadjusted $P < 0.001$, adjusted $P < 0.001$), serious / stern expression ($\chi^2(1) = 44.26$, unadjusted $P < 0.001$, adjusted $P < 0.001$), physically fit / muscular ($\chi^2(1) = 89.69$, unadjusted $P < 0.001$, adjusted $P < 0.001$), and sexualized appearance ($\chi^2(1) = 0.27$, unadjusted $P = 0.606$, adjusted $P = 1.000$).



Fig. 6: Qualitative montage of generated "trade broker" images from Lumina-Image-2.0 by nationality and three gender. All images use the prompt template "a <nationality> <gender> working as trade broker". The images depicts women across nationalities as uniformly youthful, attractive, and smiling, whereas men are older, more often serious. The nonbinary row is depicted almost entirely as young men, and only the Brazilian cell shows a visibly androgynous subject.

194 *and pretty, healthy and sporty.*” An Australian man similarly shared that, “*The typ-*
195 *ical Australian man isn’t a white rugged ‘Hugh Jackman’ type, Australian is a very*
196 *multicultural country... the imagery whitewashed that to a great extent.*”

197 These patterns often left annotators feeling excluded and a lack of belonging. An
198 American woman of color said the images, “*Didn’t look like me. It made me feel*
199 *othered*”, while a German man of immigrant background felt after seeing the images
200 “*like I am not German, or not as German [as others].*” Relatedly, a Brazilian woman
201 described feeling “*insecure*” after reviewing images. Nonbinary annotators described
202 erasure extending beyond gender. One nonbinary annotator put it: “*Non-binary people*
203 *come in all shapes, sizes, races, and levels of ability. Please teach your AI this*”, with
204 another noting the experience of viewing images as “*infuriating... They didn’t share*
205 *my age range, my body type, my level of ability... a great many of my enbie friends*
206 *were not represented among the images.*”

207 Yet, a negative effect was not universal. Counter-stereotypical representation,
208 when well executed, could challenge assumptions and feel empowering. One woman
209 reflected that “*Seeing [women] in male-dominated professions definitely challenged*
210 *me to rethink how I was looking at things.*” A Latina American noted: “*seeing Latina*
211 *women in important job positions could serve as a reminder that just because you don’t*
212 *know many women like that in your circle, it does not mean it is not attainable.*” Non-
213 binary annotators described moments of feeling affirmed by more gender-ambiguous
214 representations, even amid frustration with images overall. A nonbinary annotator
215 captured this contrast: “*Sometimes, when the images showed a male-presenting per-*
216 *son in a traditionally masculine role, or vice versa, it was upsetting; when someone*
217 *was more gender-ambiguous, it felt affirming.*”

218 Separately, we asked whether viewing the images shaped annotators’ broader sense
219 of who is expected to look a certain way or occupy certain roles. Over a quarter (27.1%)
220 reported that the images explicitly shaped their broader sense of who is expected to
221 look a certain way or occupy certain roles. Over a third (36.5%) denied being impacted,
222 sometimes attributing this to the images’ inaccuracy or lack of realism.

223 3 Discussion

224 Our findings demonstrate that T2I systems do not merely output isolated sta-
225 tistical imbalances, but reproduce a coherent set of gender logics—long theorized
226 and documented in feminist and social psychological scholarship—now encoded as
227 learned model behaviors. First, models reflect androcentrism—treating maleness as
228 the unmarked, generic human category [19, 23]. This manifests in pervasive male rep-
229 resentation defaults (Result 1) across all models, intensifying in countries with greater
230 structural gender inequality and lower income, as well as in occupations of higher
231 power. These results extend prior work on male over-representation in AI-generated
232 imagery [9, 12, 24] by showing this default holds even when prompts are balanced
233 against real-world representation, and is itself moderated by structural inequality.
234 Androcentrism in T2I models also manifests in quality discrepancies (Result 2),
235 whereby female depictions suffer systematically lower quality, particularly for women
236 in male-dominated occupations. This is consistent with evidence that treating men

as the norm can produce systems that serve women less well [25] and yield poorer outcomes for women [18, 26]. Furthermore, women from LMICs receive the lowest quality scores—in line with an intersectional account [27]. Second, models stereotype men and women in bidirectional, opposing ways (Result 3), consistent with gender polarization, whereby men and women are expected to fit distinct scripts [20]. The attribute profile observed maps onto established contrasts between agentic stereotypes attributed to men (dominance, competence, physical strength—with images of men rendered stern, fit, muscular) and communal stereotypes attributed to women (warmth, attractiveness, approachability—with images of women rendered as smiling, thin, attractive) [28, 29]. This pattern reveals a blind spot in prior work, which has focused on how women are depicted and left men’s own stereotyping largely unexamined [14, 24]. We further extend prior work documenting women’s depiction as younger than men in online media and language models [4] by showing that gender age asymmetries are part of a coherent gendered stereotype package. Finally, models exhibit nonbinary marginalization, whereby nonbinary people are not merely under-represented but poorly comprehended (Result 4), extending accounts of nonbinary erasure in classificatory AI systems [21, 22] into the generative domain.

The consistency of male defaults—alongside other gender logics—spanning different providers, architectures, and training pipelines, suggests results reflect bias embedded in the shared training data and machine learning approaches underlying current generative AI, rather than isolated design choices. Still, model variants within the same family can diverge. Flux-Schnell exhibits greater male skew than Flux for 21 of 22 nationalities evaluated. These models—while distilled from a common teacher—differ in their distillation objectives, number of sampling steps, and guidance scales. Because these factors vary simultaneously and no undistilled model from the same family is available for comparison, we cannot attribute the differences to any single factor. Nevertheless, the result has important implications: the more male-skewed model is cheaper and faster, making it more likely to be used for high-volume inference.

Annotators’ qualitative evaluations reveal that encoded gender logics have tangible impacts. Patterns of homogenization and stereotyping carry felt consequences, such as exclusion and othering. Extending evidence that counter-stereotypical T2I outputs can reduce implicit gender bias among general viewers [30], we find that members of depicted identity groups experience counter-stereotypical or gender-ambiguous representations as affirming and actively desirable when images are rendered well. These results reflect how representation can expand what someone believes is possible for themselves, or reinforce narrow, exclusionary notions of belonging and capability.

Several limitations qualify these findings. The ILO occupational data spans 57 countries across all regions globally, but reporting constraints bias the sample toward higher-income nations. Human annotation was restricted to two of the four models tested and to a subset of countries, and nonbinary annotation was further limited by recruitment constraints to three countries. Identity-matching in our design operates at the national level and cannot capture intra-national heterogeneity (e.g., caste, region, religion, class) shown to shape perceptions of representational harm [31]. T2I models also evolve rapidly, and our findings reflect specific model versions. However, as noted

above, the pattern holds across all four model providers tested, giving little basis to expect newer releases will have resolved it absent deliberate intervention.

These patterns carry stakes beyond documentation making them urgent to understand and address. Outputs can reinforce internal biases and broader societal stereotypes around how men and women are supposed to look and what spaces they are meant to occupy. Biased outputs also risk being scraped into future training data, further entrenching these patterns in subsequent models. This is consistent with a broader pattern of “encoded gender norms” documented in other AI applications, in which systems built without explicit attention to the gendered nature of data and model design reproduce existing inequalities, reinforcing gender norms as self-fulfilling prophecies [32]. Unlike more visible forms of image and media bias, where an advertisement or an editorial choice can be identified and contested, the defaults documented here operate without a single decision-maker and are repeated at a scale no prior medium can match. Compounded by a widespread belief in the neutrality of algorithmic systems, they remain largely invisible to the individuals they influence, even as they may quietly shape broader societal norms and expectations.

Addressing issues requires interventions beyond simply more women represented in training data. The biases we document could emerge from a combination of factors. Recent evidence suggests training data plays a substantial role: in CLIP and Stable Diffusion models, 60–70% of gender bias can be linearly predicted from gender co-occurrence patterns in LAION training data alone, with the remaining bias reflecting model-level amplification [11]. While training data underlying our four models is not publicly documented, the cross-model consistency we observe suggests imbalanced data is a primary, but not singular, driver. Future research should isolate the contributions of training data composition, architectural design choices, and other processes such as reinforcement learning from human feedback (RLHF) to gender bias in generation.

Our findings point to several bias mitigation directions. First, training data could be deliberately increased for underrepresented groups, with a focus on women in male-dominated occupations, people from LMICs across gender identities, and nonbinary individuals. This data should avoid replicating existing gender norms (i.e. men as fit, muscular, and stern; women as thin, young, and smiling) and instead reflect greater diversity in body size, facial expression, and age within each gender category. Such curated, diversified data could directly support mitigation interventions, such as fine-tuning. Finally, mitigation efforts should be evaluated on a comprehensive set of different dimensions related to image quality, realism, comprehension, and stereotyping utilizing annotations from people who identify with the groups being depicted. Current debiasing methods [33–35] are validated almost exclusively on the distribution of who appears, leaving open whether the resulting depictions are also realistic, dignified, and free of physical and trait-based stereotyping; while automated quality assessments remain deeply flawed (see ‘Automated quality metrics and rationale for human annotation’ in Methods). We offer our identity-matched, multi-metric annotation approach as a template for future evaluations and mitigation efforts.

4 Methods

Our methods proceeded in five stages: (1) selection of countries, occupations, and activities for prompts, (2) selection of models for analysis and image generation process, (3) human annotation and survey approach, (4) construction of the resulting datasets, and (5) our analytic approach to both the automated and human-annotation studies.

4.1 Country, occupation/activity, model selection, and image generation

4.1.1 Country selection

To ensure systematic global coverage, we began with the United Nations geoscheme (UN Statistics Division M49 standard; <https://unstats.un.org/unsd/methodology/m49/>, retrieved 22 July 2026), which divides the world into 22 subregions (e.g., Northern Africa, Western Europe, Polynesia, Southern Asia). For each of the 22 subregions, we identified the largest country by population, yielding a set of 22 countries. French Polynesia and Guam were selected as the largest entities within their respective subregions (Polynesia and Micronesia), despite their small absolute size relative to countries like the United States or China and official classifications as territories. From this set of 22 countries, a subset of 7 was selected for the human-annotation study, with one country drawn from each major world region, while also selecting countries that had sufficient representation on Prolific, the crowdworker platform that we used for human annotation recruitment. The 7 countries include: Australia, Brazil, Egypt, Germany, India, South Africa, and the United States.

4.1.2 Occupation and activity selection

We selected occupations to span high- and low-power roles, as well as male-dominated and female-dominated fields. Occupations were identified using International Labour Organization (ILO) data on female representation across jobs globally, given its cross-national representativeness. This data reflects weighted averages based on International Standard Classification of Occupations (ISCO-08) unit groups, drawn from the latest available data across 57 countries representing 24% of global employment (ILOSTAT). The 57 country sample spans diverse global regions (including Europe, Latin America, Africa, and Asia), though it skews towards high-income and upper-middle-income economies that have advanced statistical reporting capacity necessary to capture the ISCO-08 unit groups. An explicit country list is not available due to the dynamic nature of the ILOSTAT query.

We divided occupations into ten bands based on female employment ratio, in 10-percentage-point increments from 0–10% to 90–100% female-dominated, and selected six occupations from each band. Within each band, we sought to balance occupations across three power levels (low, medium, and high) targeting two occupations per level. Power level was estimated based on three factors assessed by the research team: decision-making authority, economic and social influence, and specialized expertise.

364 We note this classification is not exhaustive and involves a degree of subjectivity. Per-
365 fect balance across power levels was not achievable in all bands, as occupations in
366 the most female-dominated bands (e.g., 80–90% and 90–100% female) skewed heav-
367 ily toward low-power roles, with no high-power occupations available in these bands.
368 In these cases, we balanced the deficit by including more medium-power occupations
369 where possible, rather than defaulting entirely to low-power roles.

370 Activities were selected to reflect common, everyday activities carrying gender
371 associations, drawn primarily from the categories identified in prior research [15] on
372 gendered leisure time and occupational aspirations. For each category, we retrieved
373 the two most similar activities from an existing activity pool [36], supplementing with
374 ChatGPT-generated activities for four categories lacking close matches (build, social
375 studies, religious activities, collecting things) and retaining the category itself for three
376 categories with no meaningful variation (watch TV, dance, bike/skating). This yielded
377 45 activities spanning sports, art, gardening, games, science, pet care, handicrafts,
378 religious practice, language arts, reading, writing, music, and related domains. Of these
379 45 activities, 60% are aligned to female-coded activities, 28.9% with male-coded, and
380 11.1% with neutral (See Supplementary Note 2).

381 4.1.3 Selection of models

382 We include the following models in our analysis: Flux, Flux-Schnell, Lumina-2, and
383 Stable Diffusion 3.5 Large. These models were selected to cover a range of model
384 providers, including geographic variability. Flux and Flux-Schnell are developed by
385 the Germany-based startup BlackForestLabs, Stable Diffusion 3.5 is developed by
386 UK-based Stability AI, and Lumina 2 is developed at the Shanghai AI Laboratory.
387 Additionally, we include Flux-Schnell as a few-step diffusion model, which trades
388 performance for faster image generation.

389 We prioritize and focus on leading open-weight models (rather than closed, pro-
390 prietary models) because our broader research agenda includes bias mitigation work
391 involving fine-tuning and updating models directly. Establishing baseline bias mea-
392 surements on open-weight models allows subsequent research to fine-tune and modify
393 the same underlying models and evaluate mitigation efforts directly against this
394 paper’s findings. Furthermore, these four models see popular usage. As of August
395 2026, Flux and Flux-Schnell (Black Forest Labs) collectively average over 750,000
396 downloads per month on Hugging Face (537,643 and 221,523, respectively), Stable
397 Diffusion 3.5 Large (Stability AI) averages roughly 64,000 downloads per month,
398 and Lumina-Image 2.0 (Shanghai AI Laboratory) averages roughly 1,200 down-
399 loads per month (all download figures retrieved on 1 August 2026 from the models’
400 Hugging Face pages: <https://huggingface.co/black-forest-labs/FLUX.1-dev>, <https://huggingface.co/black-forest-labs/FLUX.1-schnell>, <https://huggingface.co/stabilityai/stable-diffusion-3.5-large>, <https://huggingface.co/Alpha-VLLM/Lumina-Image-2.0>).
401 Beyond direct downloads, these models have spawned a substantial ecosystem of
402 derivative work. Flux alone underlies over 42,000 adapters and nearly 600 fine-
403 tuned variants. The combination of direct usage and downstream development reflects
404 substantial adoption.
405
406

4.1.4 Image generation process

For each occupation/activity, we generated images using (a) gender- and nationality-neutral prompts for each occupation/activity (e.g., “an Egyptian person working as a doctor”); and (b) prompts explicitly specifying gender and nationality alongside each occupation/activity (e.g., “an Egyptian woman working as a doctor”). We generated 924,000 images from gender-neutral prompts of the form “an [nationality] person working as [occupation]” / “an [nationality] person [activity]”, across 4 text-to-image models (Flux, Flux-Schnell, Lumina2, SD3.5-Large), 22 nationalities, and ~ 105 occupations/activities. Gender was assigned by an annotation model (further discussed in “Automated classification study”) resulting in 804,258 images (87.0%) that were predicted as male or female, while 119,742 (13.0%) were identified as unknown/both/no-person. We generated 2,772,000 images that were gender and nationality specific for each occupation/activity. This included 100 images per gender, nationality, and occupation/activity prompt. These numbers are derived as follows: We have 60 occupation plus 45 activity prompts, i.e. 105 prompts. For each prompt, we generated 100 images for all combinations with 3 genders (male/female/nonbinary) and 22 nationalities across each of the 4 T2I models. From this larger pool of images, we selected 10 occupations and 10 activities for each country, balanced between male- and female-dominated roles, and restricted annotation to outputs from Flux and Stable Diffusion 3.5 Large, the two strongest-performing models in this study as only a limited number of images could be included in the survey. The image dataset thereby used for the human annotation study drew from 264,000 images.

4.2 Human annotation and survey approach

4.2.1 Recruitment, consent, and compensation

We conducted a human annotation study in which identity-matched participants rated a subset of gender- and nationality-specific images across a range of metrics. For each gender $\times$ nationality pairing, we created a separate annotation task on Prolific, restricted to participants who had identified their gender and nationality as matching that pairing in their Prolific profile. This ensured that annotators only viewed and evaluated images corresponding to their own gender and nationality. We recruited annotators who identified with the specific gender and nationality depicted in each image, based on the premise that individuals who share a given identity are better positioned to recognize stereotyping and representational harm relevant to that identity than demographically mismatched raters, while also being able to better inform how viewing AI-generated images of their particular identity group affects them. This aligns with empirical work showing that annotator identity and lived experience shape judgments of bias and harm [37], while annotators from different backgrounds notice different stereotypes in images and that outsiders to a culture (whether researchers or data annotators) often lack the cultural knowledge to recognize how cultural subject matter has been misgenerated [31].

The survey was piloted with a smaller sample prior to full deployment. Informed consent was secured from all participants prior to any research activity and was embedded at the start of the Qualtrics survey. Consent included the study’s CPHS

450 protocol ID (2025-09-18928), following the PI’s university exempt human-subjects
451 consent guidance. Because the study included participants located in the European
452 Economic Area, consent materials also included GDPR-required language. Overall,
453 170 annotators were recruited and completed the one-hour survey, receiving US\$16
454 for completing the task.

455 4.2.2 Annotation process

456 Ten annotators were recruited per gender $\times$ nationality category to evaluate images
457 matching their own gender and nationality, drawn from both gender-neutral and
458 gender- and nationality-specific prompts across the 20 occupations and activities
459 described above. Each annotator completed 20 image annotations: 2 gender-neutral
460 prompts and 18 gender- and nationality-specific prompts, randomly drawn from the
461 corresponding image pool. This yielded up to 200 annotated images per gender $\times$
462 nationality category (20 prompts, 10 images per prompt), with quotas ensuring each
463 prompt received annotations across 10 distinct images. Male and female categories
464 were annotated across all seven countries; nonbinary categories were annotated for
465 three countries (America, Australia, Germany), with a smaller number of additional
466 annotations from Brazil, as nonbinary annotation was not possible for the remaining
467 countries due to limited nonbinary participant availability on Prolific. Participants
468 rated the output on quality, naturalness, comprehension, stereotyping content, and
469 demeaning content. A closing reflection asked participants to consider how their own
470 lived experience shaped their evaluation responses. Participants were also asked how
471 viewing the generated images affected how they felt about themselves and their
472 identity group (See Supplementary Information for multi-metric annotation protocol).

473 4.3 Construction of resulting datasets

474 We constructed two datasets from the images generated for analysis. The first dataset
475 ($n = 804,258$) includes the 105 occupation/activity prompts, across 22 countries, with
476 100 outputs per occupation/activity and nationality combination across each of the
477 four models (gender-neutral); excluding images that the annotation model assigned
478 as unknown/both/no-person. This dataset was used for our automated analyses of
479 default gender representation.

480 The second dataset ($n = 3,533$ annotations) consists of human annotations drawn
481 from the gender- and nationality-specific image pool, restricted to a subset of seven
482 countries (South Africa, India, Egypt, Brazil, America, Australia, Germany), 10 occu-
483 pations and 10 activities per country, and outputs from Flux and Stable Diffusion 3.5
484 Large only (See Supplementary Table ??).

485 4.4 Analytic approach

486 4.4.1 Automated classification study

487 Using the first dataset, we assessed default gender representation by examining the
488 gender distribution of images generated from gender-neutral prompts overall, by coun-
489 try, and by model. For this, we first used YOLOv11 to detect bounding boxes around

people in all images. Then, we crop each bounding box and use Qwen3-VL-32B to predict the perceived gender of the person. Qwen3-VL can choose from four options: male, female, nonbinary, and unknown. The “unknown” label applies to images where no person is shown or there are no clear gender cues visible. Male ratios reported are the proportion of generated images that our automated classifier labels as depicting a man (see Methods), after removing those classified as unknown (see Extended Data 4). We examined cross-national variation in these defaults, testing whether default gender representation shifts as a function of country-level gender inequality (e.g., using SIGI), and specifically whether male defaults are amplified in higher-inequality contexts. We similarly examine variation by occupational power and status, testing whether default gender representation varies between high-, medium-, and low-power occupations, and whether male-default bias is more pronounced in high-status roles, paralleling but extending the finding of Guilbeault et al. [4] that age gaps were largest in high-status occupations. Finally, we conduct a cross-model comparison to assess whether these patterns are consistent across the different T2I models tested, or whether the magnitude or direction of bias varies by model.

We conducted robustness checks to confirm results in generated images. In particular to confirm whether the Chinese reversal in the trend of male default outputs was not due to errors in our automated gender-classification process, we manually reviewed a subsample of these images (400 images in total; 100 for each Chinese, Indonesian, Guam, and French Polynesia) and found the automated labels consistent with what human review would assign, suggesting the reversal reflects genuine model behavior.

4.4.2 Human-annotation study

Using the second dataset, we assessed four broad metrics. First, we examined quality through metrics of overall perceived quality, prompt comprehension, realistic scenario, and realistic person scores. We compared images of men and women, and tested whether gaps varied by country income group and by whether the depicted gender matched or violated the occupation’s gender stereotype. Second, we examined gender polarization, comparing how often annotators described men’s and women’s images with stereotyped attributes (e.g., attractive, thin, muscular, stern, sexualized, smiling). We examined skin tone, comparing average Fitzpatrick ratings across the seven annotated countries to assess whether outputs defaulted toward lighter skin tones regardless of a country’s actual racial composition. Finally, we examined nonbinary representation, comparing comprehension across male, female, and nonbinary prompts to assess erasure beyond the quality and stereotyping gaps observed between men and women.

In addition to our quantitative annotation metrics, we asked annotators several open-ended questions, including: (1) how it felt to see AI-generated images of someone sharing their own nationality and gender, and (2) whether these images shaped their sense of who is typically expected in certain occupations or activities. Responses were analyzed using an abductive thematic coding approach, combining deductive categories derived from our quantitative findings (e.g., male as the default, stereotyping, nonbinary erasure) with inductive categories emerging from responses. This yielded

372 coded excerpts across 173 annotators. To supplement responses to these two questions, we additionally reviewed annotators’ answers to other open-ended items (e.g., “do you have any other comments?”) for content that substantively addressed either of the two core questions, and incorporated relevant excerpts into our coding where applicable.

4.4.3 Automated quality metrics and rationale for human annotation

We initially conducted quality analyses on this dataset using different image reward models, specifically CLIPScore [38], ImageReward [39], and VQAScore [40]. For each generated image, these models provide a scalar score that jointly assesses general image quality and alignment with the generation prompt. However, we observed that these scores are highly divergent; consequently, image rankings as well as aggregated nationality rankings differ between reward models. More fundamentally, using such automated metrics, we cannot disentangle potential out-of-distribution effects and image failure modes. For instance, out-of-distribution effects may occur when no images of people from a certain nationality were included in the reward model’s training data. On the other hand, we cannot distinguish this from genuine failures in the image generation process, such as low image quality or inadequate representation of the prompt information. These metrics – along with related learned preference models such as PickScore [41] and HPSv2 [42] – are designed to compare models against one another rather than to compare identities within a single model and are trained on aggregated human preference data not curated or audited for demographic balance. Recent audits confirm this concern empirically. Reward-guided optimization using PickScore has been shown to disproportionately increase sexualized content for female subjects, while ImageReward-guided optimization causes generated outputs to converge toward White subjects [43]. These findings suggest that learned preference models can encode and amplify demographic biases, limiting their reliability as neutral measures of disparities across identities. Given these limitations, we relied on human annotation as a primary method for quality assessment.

Our multi-metric annotation design contributes valuable methodological insights largely absent from work in this space. Prior work has typically measured gender bias through a single type of evidence, such as representation counts from neutral prompts [5, 7, 9], embedding-based similarity to a gendered default [24], or trait associations inferred from demographic output [44]. Missing from the existing literature are assessments of image quality or realism across different demographics, or grounding stereotype measurement in established gender research. We introduce a multi-metric annotation protocol that captures comprehension, quality, naturalness, and gender stereotyping content including physical-appearance and trait-based attributes grounded in established gender research, rated by annotators that match the identity depicted in the image. This design allows us to distinguish between different failure modes. For example, an image may be perceived as having high comprehension, but the person in the image could be unrealistic or reflect particular gender stereotypes. Our multi-metric annotation protocol and identity-matched annotated dataset can serve as a template for future gender bias evaluations, extending prior automated

577 and human-annotation approaches that have typically relied on a single dimension or
578 lacked identity-matched recruitment.

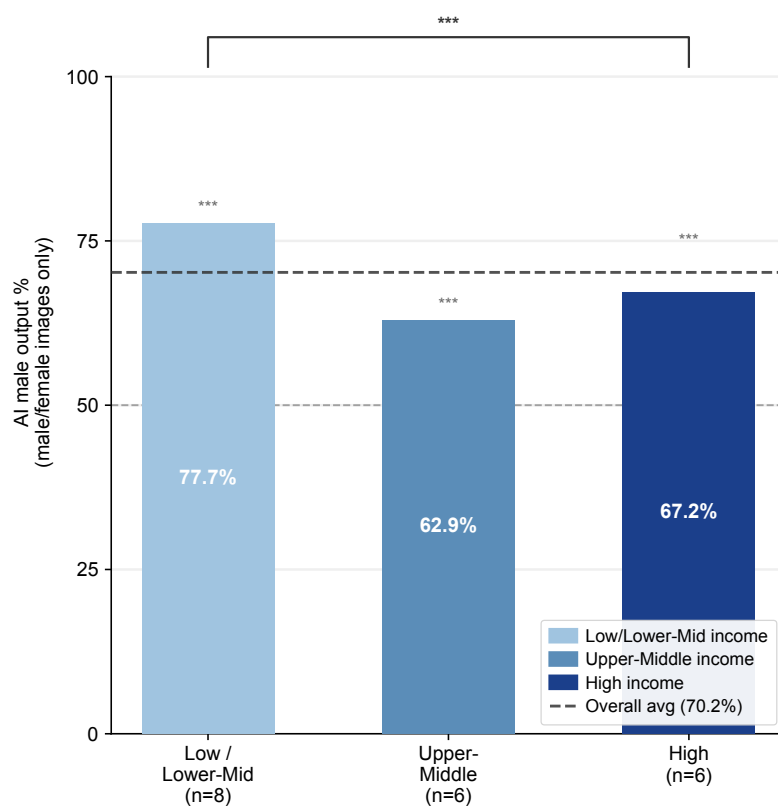


Fig. Extended Data 1: Male default, by country income level. Bars show mean percent male AI-generated output for each World Bank income group, averaged across all models, prompts, and datasets (20 of 22 nationalities; Guam and French Polynesia are excluded as territories without independent World Bank income classifications; $n = 297,559 / 214,537 / 219,920$ images for Low/Lower-Middle, Upper-Middle, and High income groups, respectively). Dark dashed line, overall average (70.2%); light dashed line, 50% parity. Asterisks denote significance vs. 50% parity (one-sample t -test) and between Low/Lower-Mid and High income groups (two-sample t -test): *** $P < 0.0001$.

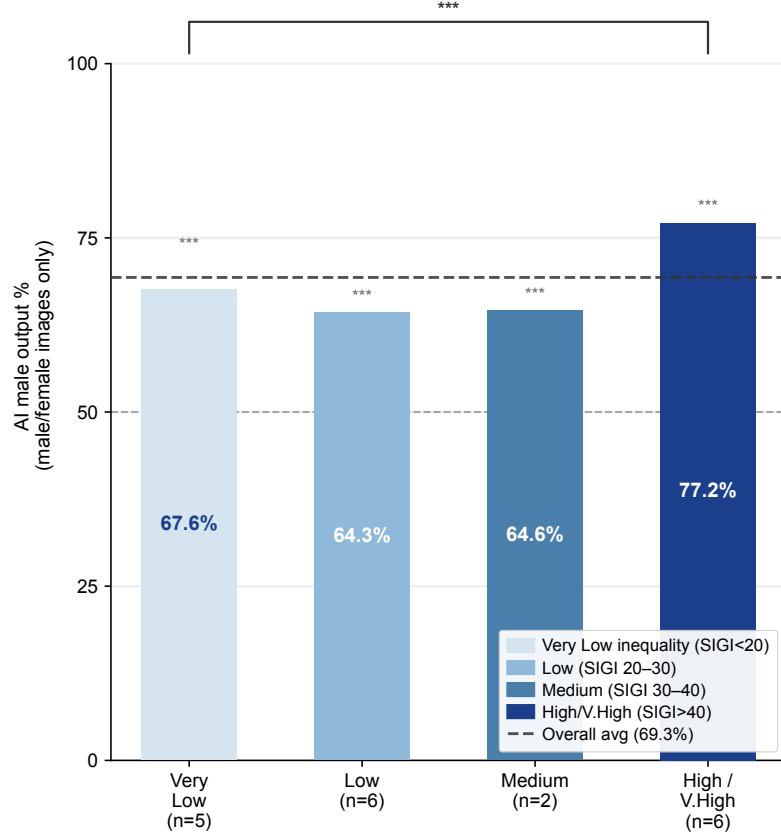


Fig. Extended Data 2: Male default, by levels of gender inequality (SIGI). Bars show mean percent male AI-generated output for each SIGI gender-inequality group, averaged across all models, prompts, and datasets (19 of 22 nationalities; Uzbekistan, Guam, and French Polynesia are excluded as they lack SIGI scores; $n = 183,471 / 219,796 / 68,192 / 222,649$ images for Very Low, Low, Medium, and High/Very High inequality groups, respectively). Dark dashed line, overall average (69.3%); light dashed line, 50% parity. Asterisks denote significance vs. 50% parity (one-sample t -test) and between Very Low and High/Very High inequality groups (two-sample t -test): *** $P < 0.0001$.

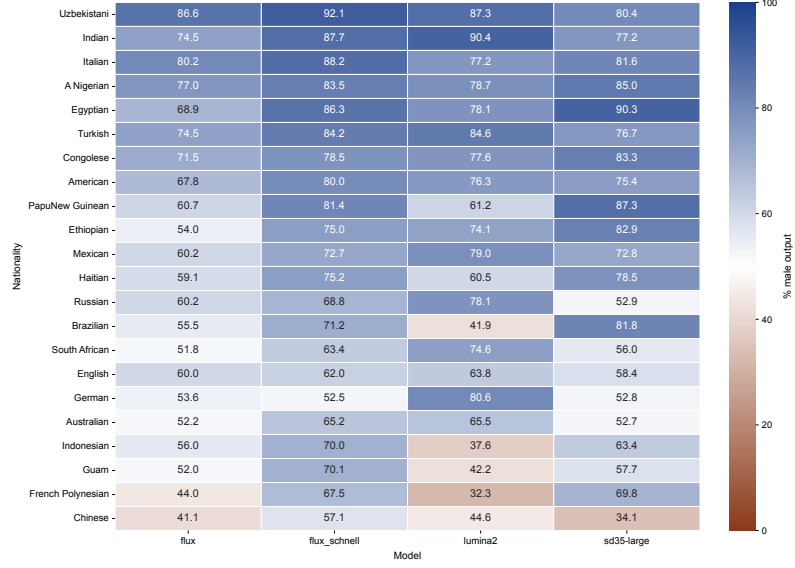


Fig. Extended Data 3: AI male output by nationality and model. Heatmap cells show mean percent male AI-generated output for each of 22 nationalities (rows, sorted by overall percent male output, descending) across 4 text-to-image models (columns: flux, flux_schnell, lumina2, sd35-large; $N = 804,258$ images total, 6,690–10,033 per nationality-model cell). Color scale is a diverging gradient from female-associated orange through white at 50% parity to male-associated blue, using the same male/female color convention applied throughout the paper. Rows are ordered by each nationality’s overall mean percent male output, averaged across all 4 models and prompts).

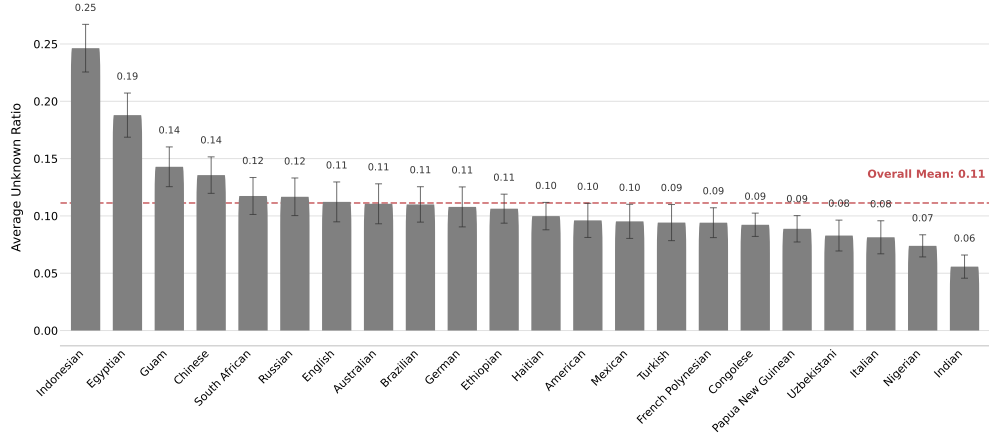


Fig. Extended Data 4: Proportion of images assigned an unknown perceived gender label across nationalities. Proportion of generated images resulting from gender-neutral occupational prompts where shown individuals could not be reliably classified into binary perceived gender categories and were assigned an unknown label (grey bars), shown across 22 evaluated nationalities. Bar heights represent the sample mean proportion across all pooled text-to-image models and prompts; error bars represent 95% confidence intervals ($\pm 1.96 \times \text{s.e.m.}$). Numerical values above each error bar denote the mean unknown proportion for that nationality. The horizontal dashed line shows the overall mean across all evaluated nationalities (mean = 11.1%, standard deviation s.d. = 16.3%, across $N = 9,240$ total nationality-prompt evaluations; $n = 420$ to 420 evaluations per nationality). Mean unknown proportions range from a minimum of Indian (5.6%) and Nigerian (7.4%) to a maximum of Egyptian (18.8%) and Indonesian (24.6%).

Occupation	Power level	ILO real-world male (% band)	Actual male (%)	AI predicted male (%)	Difference (pp)	Direction
archivist	Low	10-20	14.7	75.2	60.5	More male
meteorologist	Medium	40-50	42.7	95.3	52.6	More male
psychologist	Medium	10-20	16.4	66.5	50.2	More male
cleaner	Low	30-40	32.8	78.5	45.7	More male
translator	Low	30-40	32.1	75.6	43.5	More male
financial advisor	High	40-50	45.2	86.8	41.7	More male
political scientist	High	30-40	39.9	80.3	40.4	More male
optometrist	Medium	30-40	30.7	70.4	39.6	More male
shopkeeper	Low	50-60	53.6	88.8	35.2	More male
chef	Medium	60-70	63.1	98.1	35.0	More male
judge	High	50-60	52.6	87.5	34.9	More male
physiotherapist	Medium	30-40	30.8	64.6	33.8	More male
advertising manager	High	50-60	53.0	86.0	32.9	More male
accountant	Medium	50-60	51.0	82.8	31.8	More male
choreographer	Low	20-30	22.3	50.4	28.0	More male
butcher	Low	70-80	72.8	99.3	26.4	More male
crop farm laborer	Low	60-70	67.5	93.9	26.4	More male
economist	High	60-70	67.2	92.3	25.1	More male
dentist	Medium	40-50	44.9	69.0	24.0	More male
architect	Medium	70-80	73.7	97.6	23.8	More male
nutritionist	Medium	10-20	17.7	40.5	22.8	More male
journalist	Medium	60-70	66.5	89.0	22.5	More male
author	Medium	50-60	56.9	79.2	22.2	More male
lawyer	High	60-70	68.5	90.8	22.2	More male
domestic cleaner	Low	0-10	6.3	27.7	21.4	More male
senior government official	High	70-80	71.5	90.8	19.3	More male
crop farmer	Low	70-80	74.9	92.5	17.6	More male
secretary	Low	10-20	17.0	33.7	16.6	More male
photographer	Low	80-90	80.7	97.1	16.4	More male
web developer	Medium	80-90	82.4	97.7	15.2	More male
sports coach	Low	80-90	83.4	97.8	14.4	More male
chief executive	High	70-80	79.5	93.7	14.2	More male
travel guide	Low	60-70	63.3	75.9	12.6	More male
trade broker	High	80-90	86.6	97.6	11.0	More male
security guard	Low	90-100	91.0	99.0	8.1	More male
librarian	Medium	30-40	31.9	38.8	6.9	More male
therapist	Medium	10-20	29.5	36.4	6.9	More male
firefighter	Medium	90-100	90.2	97.1	6.8	More male
head of village	High	80-90	84.8	91.5	6.7	More male
physicist	High	90-100	90.0	95.6	5.5	More male
personal care worker	Low	0-10	9.9	15.1	5.1	More male
mechanical engineer	High	90-100	95.5	99.9	4.4	More male
police officer	Medium	80-90	87.6	92.0	4.3	More male
professional athlete	Medium	90-100	93.0	95.6	2.7	More male
carpenter	Low	90-100	98.1	100.0	1.8	More male
childcare worker	Low	0-10	2.6	1.5	-1.0	More female
fitness instructor	Low	50-60	59.0	57.5	-1.6	More female
midwifery professional	Low	0-10	4.4	0.4	-4.0	More female
visual artist	Medium	70-80	70.9	65.9	-5.0	More female
early childhood educator	Low	0-10	7.1	1.2	-6.0	More female
child care services manager	Medium	0-10	9.7	2.8	-6.9	More female
receptionist	Low	20-30	22.1	14.1	-8.0	More female
hairstylist	Low	40-50	44.5	34.8	-9.7	More female
social work professional	Medium	20-30	29.5	19.5	-9.9	More female
community health worker	Medium	20-30	26.7	15.6	-11.1	More female
teaching professional	Medium	40-50	43.7	31.7	-12.0	More female
administrative secretary	Low	40-50	41.2	28.5	-12.7	More female
beautician	Low	10-20	15.5	2.3	-13.2	More female
fashion model	Low	20-30	20.7	6.6	-14.0	More female
nurse	Medium	20-30	23.2	3.9	-19.3	More female

Table Extended Data 1: Direction and magnitude of the AI gender shift for all 60 occupations. Rows are occupations, ordered by the difference between AI-predicted and real-world male share. Power level (high, medium, low) is as defined in Methods. ILO real-world male (% band) is the 10-percentage-point range containing the occupation’s male share of the workforce. Actual male (%) is 100 minus the actual female share from ILO data, and is the real-world value used for all comparisons. AI predicted male (%) is the mean across 22 nationalities and 4 text-to-image models, from male- or female-classified images per occupation (median). Values are means. Difference is in percentage points (AI predicted male minus actual male). Direction is labeled as more male for a positive difference and as more female for a negative one; 45 occupations are more male and 15 more female.

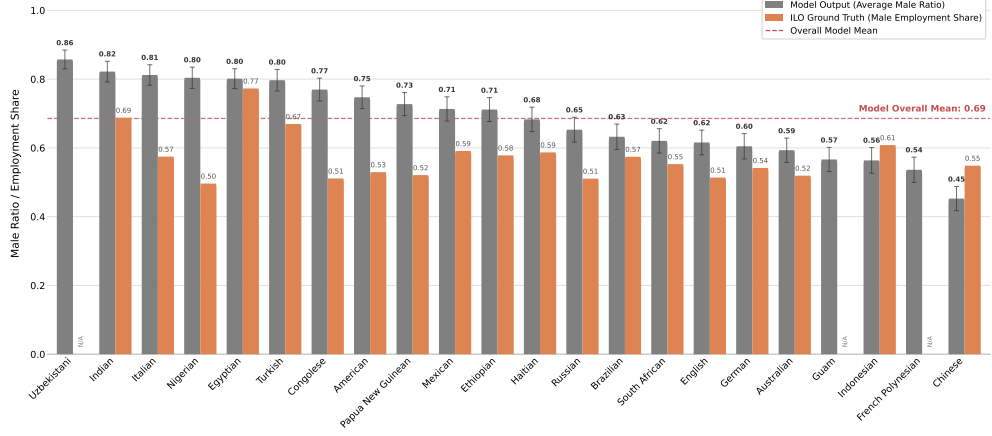


Fig. Extended Data 5: Comparison of model-generated perceived male depiction ratios against national workforce employment statistics. Male ratio in text-to-image generations produced from gender-neutral occupational prompts (grey bars) contrasted with national male labor force participation shares from International Labour Organization (ILOSTAT) records (orange bars) across 22 nationalities. For model generations, bar heights denote sample means pooled across all evaluated text-to-image models and prompt variations; error bars indicate 95% confidence intervals ($\pm 1.96 \times \text{s.e.m.}$). Numerical values next to each bar denote the observed proportion (rounded to two decimal places). The dashed horizontal line indicates the overall mean of male images across all evaluated model prompts (mean = 0.69, standard deviation s.d. = 0.37, across $N = 9,240$ evaluations; $n = 420$ evaluations per nationality). National workforce statistics were available for $k = 19$ of the 22 nationalities and no ILO records were available for Uzbekistan, Guam, or French Polynesia. Across the 19 matched nationalities, generated images substantially overrepresented men relative to empirical workforce shares (mean model share = 0.69 vs. mean ILO share = 0.57; two-sided paired t -test: $t(18) = 5.12$, $P < 0.001$; two-sided Wilcoxon signed-rank test: $W = 9.0$, $P < 0.001$).

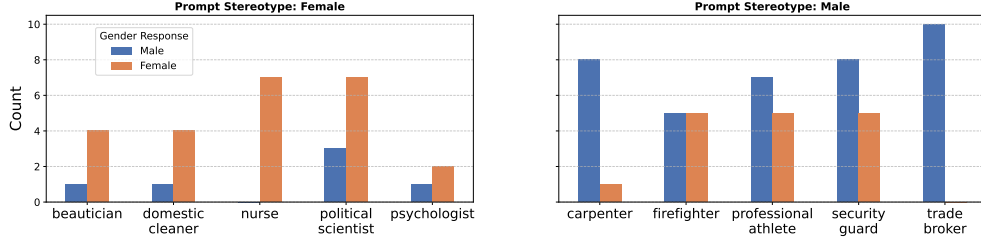


Fig. Extended Data 6: Perceived binary gender attributions for non-binary prompts across stereotyped occupations. Frequency of annotator perceived gender labels for images generated from non-binary prompts across female-stereotyped occupations (**a**, left) and male-stereotyped occupations (**b**, right). Bar heights indicate absolute label counts categorized as male (blue bars) or female (orange bars). Across all non-binary prompt evaluations ($N = 279$ total annotations), depicted subjects were mostly classified as non-binary or unclear ($n = 195$ annotations, 69.9%), while binary gender was perceived in $n = 84$ instances (30.1%). Among binary classifications, perceived gender strongly aligned with the stereotypical gender associated with the prompted occupation: in female-stereotyped occupations (beautician, domestic cleaner, nurse, political scientist, psychologist; total $n = 30$ binary annotations), 24 were perceived as female (80.0%) and 6 as male (20.0%); in male-stereotyped occupations (carpenter, firefighter, professional athlete, security guard, trade broker; total $n = 54$ binary annotations), 38 were perceived as male (70.4%) and 16 as female (29.6%). The association between occupational stereotype and perceived binary gender direction was statistically significant (two-sided Pearson’s χ^2 test of independence: $\chi^2(1) = 19.62$, $P < 0.001$; two-sided Fisher’s exact test: odds ratio OR = 9.50, 95% confidence interval [3.26, 27.65], $P < 0.001$).

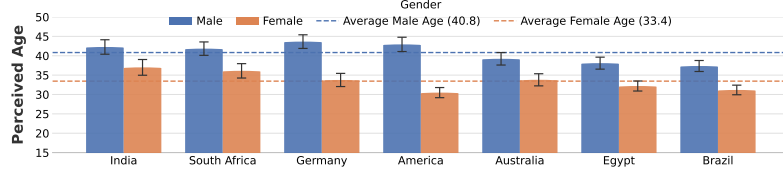


Fig. Extended Data 7: Perceived age differences in generated individuals across countries and gender. Mean perceived age of generated individuals in occupational depictions, grouped by nationality and gender (blue bars, men; orange bars, women). Perceived age ranges assigned by human annotators were converted to scalar midpoint values (18–24 years as 21.0, 25–34 years as 29.5, 35–44 years as 39.5, 45–54 years as 49.5, 55–64 years as 59.5, and ≥ 65 years as 69.5). Bar heights indicate sample means; error bars denote 95% confidence intervals from Student’s t -distribution critical values ($\pm t_{0.975, \text{d.f.}} \times \text{s.e.m.}$). The dashed horizontal lines denote the overall mean perceived age across all evaluated countries for men (dashed blue line; mean = 40.8 years, standard deviation s.d. = 12.6, $n = 1,449$) and women (dashed orange line; mean = 33.4 years, standard deviation s.d. = 11.7, $n = 1,442$; total $N = 2,891$ evaluations; $n = 180$ to 243 evaluations per country–gender group). Across all countries, women were perceived as substantially younger than men (mean difference = 7.4 years; two-sided Welch’s t -test: $t(2874.5) = 16.32$, $P < 0.001$; two-way ANOVA main effect of gender: $F(1, 2,877) = 271.53$, $P < 0.001$; main effect of country: $F(6, 2,877) = 11.92$, $P < 0.001$; gender $\times$ country interaction: $F(6, 2,877) = 5.75$, $P < 0.001$).

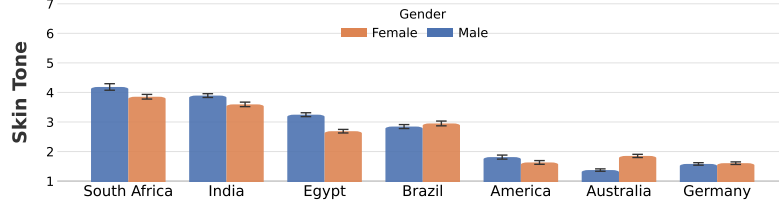


Fig. Extended Data 8: Perceived skin tone ratings across nationalities and gender. Mean perceived skin tone ratings of persons in generated images across all 7 countries, grouped by prompt gender (blue bars, men; orange bars, women). Perceived skin tone was evaluated on the Fitzpatrick scale (integer ratings from 1 to 7, where higher scores represent darker perceived skin tones). Bar heights represent sample means; error bars denote standard error of the mean ($\pm 1 \times \text{s.e.m.}$). Across all evaluated images ($N = 2,855$ total evaluations; $n = 1,433$ men, $n = 1,422$ women; $n = 179$ to 239 per country–gender subgroup), perceived skin tone varied substantially by national context (overall mean across countries: men = 2.65 ± 0.04 s.e.m., women = 2.61 ± 0.03 s.e.m.). A two-way ANOVA shows significant differences across countries ($F(6, 2,841) = 445.71$, $P < 0.001$) and a significant main effect of gender ($F(1, 2,841) = 8.16$, $P = 0.004$), as well as a significant country $\times$ gender interaction ($F(6, 2,841) = 12.36$, $P < 0.001$), reflecting higher perceived skin tone ratings for men compared to women in several countries (for example, Egypt: male mean = 3.25 vs. female mean = 2.69 ; two-sided Welch’s t -test: $t(384.5) = 6.31$, $P < 0.001$) but reversed or small differences in others.

References

- [1] Adobe: [Media Alert] Inaugural Adobe Creators' Toolkit Report: 86 Percent of Global Creators Use Creative Generative AI, See It Boosting Creator Economy. Adobe Newsroom. Accessed: 2026-08-21. <https://news.adobe.com/news/2025/10/adobe-max-2025-creators-survey>
- [2] Valyaeva, A.: People Are Creating an Average of 34 Million Images Per Day. Statistics for 2024. Everypixel. Everypixel Journal. Accessed: 2026-09-11 (2023)
- [3] Santoniccolo, F., Trombetta, T., Paradiso, M.N., Rollè, L.: Gender and media representations: A review of the literature on gender stereotypes, objectification and sexualization. *International Journal of Environmental Research and Public Health* **20**(10), 5770 (2023) <https://doi.org/10.3390/ijerph20105770>
- [4] Guilbeault, D., Delecourt, S., Desikan, B.S.: Age and gender distortion in online media and large language models. *Nature* **646**(8087), 1129–1137 (2025) <https://doi.org/10.1038/s41586-025-09581-z>
- [5] Cho, J., Zala, A., Bansal, M.: 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 (ICCV)*, pp. 3020–3031 (2023)
- [6] Wan, Y., Subramonian, A., Ovalle, A., Lin, Z., Suvarna, A., Chance, C., Bansal, H., Pattichis, R., Chang, K.-W.: Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030* (2024)
- [7] Luccioni, S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Evaluating societal representations in diffusion models. In: *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, vol. 36, pp. 56338–56351 (2023)
- [8] Naik, R., Nushi, B.: Social biases through the text-to-image generation lens. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 786–808 (2023)
- [9] Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 1493–1504 (2023). <https://doi.org/10.1145/3593013.3594095>
- [10] Sun, L., Wei, M., Sun, Y., Suh, Y.J., Shen, L., Yang, S.: Smiling women pitching down: auditing representational and presentational gender biases in image-generative AI. *Journal of Computer-Mediated Communication* **29**(1), 045 (2024) <https://doi.org/10.1093/jcmc/zmad045>

- [11] Gırrbach, L., Alaniz, S., Smith, G., Akata, Z.: A large scale analysis of gender biases in text-to-image generative models. arXiv preprint arXiv:2503.23398 (2025)
- [12] AlDahoul, N., Rahwan, T., Zaki, Y.: AI-generated faces influence gender stereotypes and racial homogenization. *Scientific Reports* **15**(1), 14449 (2025) <https://doi.org/10.1038/s41598-025-99623-3>
- [13] Jha, A., Prabhakaran, V., Denton, R., Laszlo, S., Dave, S., Qadri, R., Reddy, C.K., Dev, S.: 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 (ACL)* (2024)
- [14] Ghosh, S., Venkit, P.N., Gautam, S., Wilson, S., Caliskan, A.: Do generative AI models output harm while representing non-western cultures: Evidence from a community-centered approach. In: *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (2024)
- [15] Lee, B., Skinner, O.D., McHale, S.M.: Links between gendered leisure time in childhood and adolescence and gendered occupational aspirations. *Journal of Adolescence* **62**, 96–107 (2018) <https://doi.org/10.1016/j.adolescence.2017.10.011>
- [16] Yakefu, A., Apolinário, Solaiman, I.: State of Open Models: Summer 2026 Observations. Hugging Face Blog (2026). <https://huggingface.co/blog/state-of-open-models-summer-2026>
- [17] Mozur, P., Satariano, A., Krolik, A.: China’s A.I. Is Surging Across Africa. That Should Worry Silicon Valley. *The New York Times*. <https://www.nytimes.com/2026/08/05/technology/ai-china-africa.html>. Accessed: 2026-10-05 (2026)
- [18] Bailey, A.H., LaFrance, M., Dovidio, J.F.: Is man the measure of all things? a social cognitive account of androcentrism. *Personality and Social Psychology Review* **23**(4), 307–331 (2019) <https://doi.org/10.1177/1088868318782848>
- [19] Gilman, C.P.: *The Man-Made World; Or, Our Androcentric Culture*. T. Fisher Unwin, London (1911)
- [20] Bem, S.L.: *The Lenses of Gender: Transforming the Debate on Sexual Inequality*. Yale University Press, New Haven (1993)
- [21] Keyes, O.: The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* **2**(CSCW), 88–18822 (2018) <https://doi.org/10.1145/3274357>
- [22] Scheuerman, M.K., Paul, J.M., Brubaker, J.R.: How computers see gender: An

- 651 evaluation of gender classification in commercial facial analysis services. Proceed-
652 ings of the ACM on Human-Computer Interaction **3**(CSCW), 144–114433 (2019)
653 <https://doi.org/10.1145/3359246>
- 654 [23] Hegarty, P., Buechel, C.: Androcentric reporting of gender differences in APA
655 journals: 1965–2004. Review of General Psychology **10**(4), 377–389 (2006) <https://doi.org/10.1037/1089-2680.10.4.377>
656
- 657 [24] Ghosh, S., Caliskan, A.: ‘person’ == light-skinned, western man, and sexual-
658 ization of women of color: Stereotypes in stable diffusion. In: Findings of the
659 Association for Computational Linguistics: EMNLP 2023 (2023)
- 660 [25] Criado Perez, C.: Invisible Women: Data Bias in a World Designed for Men.
661 Abrams Press, New York (2019)
- 662 [26] Beery, T.A.: Gender bias in the diagnosis and treatment of coronary artery dis-
663 ease. Heart & Lung: The Journal of Acute and Critical Care **24**(6), 427–435
664 (1995) [https://doi.org/10.1016/s0147-9563\(95\)80020-4](https://doi.org/10.1016/s0147-9563(95)80020-4)
- 665 [27] Crenshaw, K.: Demarginalizing the intersection of race and sex: A black feminist
666 critique of antidiscrimination doctrine, feminist theory and antiracist politics.
667 University of Chicago Legal Forum **1989**(1), 139–167 (1989)
- 668 [28] Bakan, D.: The Duality of Human Existence: An Essay on Psychology and
669 Religion. Rand McNally, Chicago (1966)
- 670 [29] Eagly, A.H., Steffen, V.J.: Gender stereotypes stem from the distribution of
671 women and men into social roles. Journal of Personality and Social Psychology
672 **46**(4), 735–754 (1984) <https://doi.org/10.1037/0022-3514.46.4.735>
- 673 [30] Sim, M., Brigham, N.G., Kohno, T., Charlesworth, T.E.S., Caliskan, A.: Biased ai
674 outputs can impact humans’ implicit bias: A case study of the impact of gender-
675 biased text-to-image generators. In: Proceedings of the Eighth AAAI/ACM
676 Conference on AI, Ethics, and Society (AIES-25), vol. 8, pp. 2375–2386 (2025).
677 <https://doi.org/10.1609/aies.v8i3.36723> . <https://doi.org/10.1609/aies.v8i3.36723>
- 678 [31] Qadri, R., Shelby, R., Bennett, C.L., Denton, E.: AI’s regimes of representation:
679 A community-centered study of text-to-image models in south asia. In: Proceed-
680 ings of the 2023 ACM Conference on Fairness, Accountability, and Transparency
681 (FAccT), pp. 506–517 (2023)
- 682 [32] Smith, G.: Mindsets and management: AI and gender (in)equitable access to
683 finance. arXiv preprint arXiv:2504.07312 (2025)
- 684 [33] Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni,
685 S., Kersting, K.: Fair diffusion: Instructing text-to-image generation models on
686 fairness. arXiv preprint arXiv:2302.10893 (2023)

- 687 [34] Shen, X., Du, C., Pang, T., Lin, M., Wong, Y., Kankanhalli, M.: Finetuning text-
688 to-image diffusion models for fairness. In: The Twelfth International Conference
689 on Learning Representations (ICLR) (2024)
- 690 [35] Parihar, R., Bhat, A., Basu, A., Mallick, S., Kundu, J.N., Babu, R.V.: Balancing
691 act: Distribution-guided debiasing in diffusion models. In: Proceedings of the
692 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
693 pp. 6668–6678 (2024)
- 694 [36] Wilson, S.R., Mihalcea, R.: Measuring semantic relations between human activ-
695 ities. In: Proceedings of the Eighth International Joint Conference on Natural
696 Language Processing (IJCNLP) (2017)
- 697 [37] Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., Smith, N.A.: Anno-
698 tators with attitudes: How annotator beliefs and identities bias toxic language
699 detection. In: Proceedings of the 2022 Conference of the North American Chapter
700 of the Association for Computational Linguistics (NAACL) (2022)
- 701 [38] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A
702 reference-free evaluation metric for image captioning. In: Proceedings of the 2021
703 Conference on Empirical Methods in Natural Language Processing (EMNLP)
704 (2021)
- 705 [39] Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: ImageRe-
706 ward: Learning and evaluating human preferences for text-to-image generation.
707 In: Advances in Neural Information Processing Systems (NeurIPS), vol. 36 (2023)
- 708 [40] Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.:
709 Evaluating text-to-visual generation with image-to-text generation. In: European
710 Conference on Computer Vision (ECCV) (2024)
- 711 [41] Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-Pic:
712 An open dataset of user preferences for text-to-image generation. In: Advances
713 in Neural Information Processing Systems (NeurIPS), vol. 36 (2023)
- 714 [42] Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference
715 score v2: A solid benchmark for evaluating human preferences of text-to-image
716 synthesis. arXiv preprint arXiv:2306.09341 (2023)
- 717 [43] Abdel Magid, S., Guo, G., Tureci, E., Dharmasiri, A., Ramaswamy, V.V., Pfister,
718 H., Russakovsky, O.: Bias at the end of the score. arXiv preprint arXiv:2604.13305
719 (2026)
- 720 [44] Fraser, K.C., Kiritchenko, S., Nejadgholi, I.: A friendly face: Do text-to-image
721 systems rely on stereotypes when the input is under-specified? arXiv preprint
722 arXiv:2302.07159 (2023)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.pdf](#)