Race and Gender in LLM-Generated Personas: A Large-Scale Audit of 41 Occupations

ILONA VAN DER LINDEN, Santa Clara University, USA
SAHANA KUMAR, Santa Clara University, USA
ARNAV DIXIT, Santa Clara University, USA
AADI SUDAN, Santa Clara University, USA
SMRUTHI DANDA, Santa Clara University, USA
DAVID C. ANASTASIU, Santa Clara University, USA
KAI LUKOFF, Santa Clara University, USA

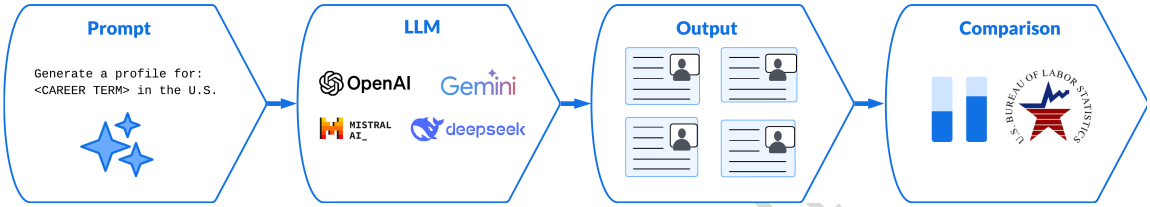

Fig. 1. Overview of our study pipeline: We prompt four large language models (OpenAI, Gemini, Mistral, DeepSeek) to generate structured occupational personas (e.g., name, age, gender, ethnicity, motivations, biography), then compare the demographic distributions in these outputs against U.S. Bureau of Labor Statistics (BLS) data.

Generative AI tools are increasingly used to create portrayals of people in occupations, raising concerns about how race and gender are represented. We conducted a large-scale audit of over 1.5 million occupational personas across 41 U.S. occupations, generated by four large language models with different AI safety commitments and countries of origin (U.S., China, France). Compared with Bureau of Labor Statistics data, we find two recurring patterns: systematic shifts, where some groups are consistently under- or overrepresented, and stereotype exaggeration, where existing demographic skews are amplified. On average, White (−31pp) and Black (−9pp) workers are underrepresented, while Hispanic (+17pp) and Asian (+12pp) workers are overrepresented. These distortions can be extreme: for example, across all four models, Housekeepers are portrayed as nearly 100% Hispanic, while Black workers are erased from many occupations. For HCI, these findings show provider choice materially changes who is visible, motivating model-specific audits and accountable design practices.

Authors' Contact Information: Ilona van der Linden, Santa Clara University, Santa Clara, California, USA, lonavdlin@gmail.com; Sahana Kumar, Santa Clara University, Santa Clara, California, USA, skumar9@scu.edu; Arnav Dixit, Santa Clara University, Santa Clara, California, USA, adixit3@scu.edu; Aadi Sudan, Santa Clara University, Santa Clara, California, USA, asudan@scu.edu; Smruthi Danda, Santa Clara University, Santa Clara, California, USA, sdanda@scu.edu; David C. Anastasiu, Santa Clara University, Santa Clara, California, USA, danastasiu@scu.edu; Kai Lukoff, Santa Clara University, Santa Clara, California, USA, klukoff@scu.edu.

CCS Concepts: • **Human-centered computing -> Empirical studies in HCI**; • **Human-centered computing -> HCI design and evaluation methods**; • **Computing methodologies -> Natural language generation**; • **Information systems -> Content analysis and representation**; • **Social and professional topics -> Ethics, social responsibility, and policy**;

Additional Key Words and Phrases: representational bias, generative AI, persona generation, occupational stereotypes, fairness, human–computer interaction

ACM Reference Format:

Ilona van der Linden, Sahana Kumar, Arnav Dixit, Aadi Sudan, Smruthi Danda, David C. Anastasiu, and Kai Lukoff. 2018. Race and Gender in LLM-Generated Personas: A Large-Scale Audit of 41 Occupations. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 29 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Technology services increasingly search, filter, and generate media in which people are represented by race and gender. A common and well-documented concern is the underrepresentation of minority groups in professional roles. For example, television and film have historically depicted men in a much greater diversity of occupations than women [63], a study of digital platforms such as Twitter, Wikipedia, and Shutterstock found that they disproportionately present librarians and nurses as women and programmers and engineers as men [64], and Google Image Search results systematically underrepresent women across a variety of US occupations [41, 50]. Research in psychology and education demonstrates the consequences: when young people do not "see themselves" represented, they are less likely to envision certain careers as possible or desirable [46, 54], more vulnerable to stereotype threat in performance domains [66], and conversely, more motivated and confident when exposed to counter-stereotypical role models [18, 20, 67]. In response, many organizations now make deliberate efforts to feature underrepresented groups in advertising, promotional materials, and media campaigns.

However, problems also arise in the opposite direction, through overrepresentation or misleading representation. In February 2025, Google's newly released Gemini AI image generator produced images of women and people of color as Nazi soldiers. In response to widespread criticism, Google admitted that it had 'missed the mark' and suspended the tool for several months [33]. More subtle cases include images that depict minorities in plausible but highly unlikely scenarios, raising concerns about "diversity-washing," the strategic overrepresentation of minority groups to signal inclusivity without reflecting real-world proportions or experiences [6]. For instance, an AI image generator might produce a boardroom photo in which nearly all executives are women of color, a scenario that is theoretically possible but statistically very rare in today's global workforce. These cases demonstrate that fair representation in computer-generated media is not simply a matter of maximizing diversity, but rather a complex sociotechnical challenge requiring careful analysis.

These issues matter not only for images, but also for text. LLM-generated personas are increasingly used as infrastructure in domains: in UX and design research [59], in industry to support product development [44], and in creative writing and media [47]. In these contexts, personas are not decorative, but shape how organizations conceptualize users, how fictional worlds are populated, and how datasets represent social groups. Building on prior auditing methods that compare generated media with real-world benchmarks, we apply this approach to LLM-generated occupational personas at scale across multiple models. Specifically, we generated over 1.5 million profiles across 41 U.S. occupations using four prominent language models (GPT-4, Gemini 2.5, DeepSeek V3.1, and Mistral-medium). We compared their demographic distributions with data from the U.S. Bureau of Labor Statistics (BLS).

We seek to answer the following research questions:

- **RQ1:** How do race and gender representations in AI-generated personas systematically differ from U.S. demographic statistics? To what extent do models exhibit average bias across occupations?
- **RQ2:** To what extent do AI-generated personas deviate from reality in systematic ways? Where do we see stereotype exaggeration, with models amplifying existing demographic majorities?
- **RQ3:** How do systematic representational biases differ across large language models, particularly those with different safety commitments and countries of origin?

Together, these questions examine not only whether representational biases exist (RQ1), but also how they systematically take shape (RQ2), and whether they differ across models with distinct institutional and cultural origins (RQ3).

We find that large language models introduce patterned distortions into occupational personas. Hispanic and Asian workers are often overrepresented, while Black and White workers are consistently suppressed. At the extremes, Housekeepers are portrayed as overwhelmingly Hispanic and female across all four models, while Black workers are nearly absent from many occupations outside a few stereotyped roles such as Security Guard or Bus Driver. While the precise offsets vary between providers, these findings suggest that representational bias in LLMs is structured rather than incidental.

Based on these results, we then turn to design implications for HCI. Rather than leaving representation as an incidental artifact, persona systems could make it an explicit design dimension. Interfaces might, for instance, show users how outputs diverge from demographic baselines or generate a small set of personas to highlight diversity, instead of returning a single default profile. When paired with model-specific audits, such mechanisms could advance practical ways of making representational choices visible and governable in practice.

2 Motivations

2.1 Algorithmic Bias in Representational Systems

Algorithmic systems are often wrapped in claims of neutrality, yet research across HCI and adjacent fields has long shown that they can reproduce and even amplify social inequities. Early conceptual work identified multiple pathways through which bias enters technical systems: social and historical patterns embedded in data, technical design decisions, and deployment contexts [27]. Building on this foundation, empirical audits have shown how algorithmic outputs can distort the visibility of social groups and shape public perceptions.

Search engines are a well-documented case. Google Image Search systematically underrepresents women in occupational queries and exaggerates stereotypical associations, shifting participants' beliefs about gender composition in those fields [41]. Racial disparities are similarly evident: image search underrepresents racial minorities, with measurable effects on users' feelings of belonging in professional domains [50]. Large-scale analyses across platforms such as Google and Wikipedia further show that visual content amplifies gender bias more strongly than text, and that this amplification shapes public beliefs about social groups [35]. Collectively, this body of work makes clear that algorithmic bias is not simply a technical artifact, but a form of representational harm with consequences for identity, aspiration, and opportunity.

2.2 Representational Bias in Text-to-Image Generation

Recent advances in text-to-image (T2I) systems such as Stable Diffusion, DALL·E, and Midjourney have expanded concerns about representational bias beyond search into synthetic media. These models consistently amplify demographic stereotypes at scale. Prompts for occupations or traits often yield images that overrepresent men and White individuals, while associating marginalized groups with criminality, poverty, or hypersexualization [12]. A survey of the literature confirms that T2I systems routinely display gender, skin tone, and geocultural biases, and that current mitigation strategies remain limited [71].

Domain-specific audits underscore the stakes. AI-generated depictions of surgeons and physicians disproportionately portray them as White men, diverging sharply from U.S. workforce demographics [1, 43]. In multiple professional domains, men appear in more than three-quarters of images of lawyers, doctors, engineers, and scientists [32]. Beyond occupations, *adultification bias* has been observed, with Black girls depicted as older and more sexualized than their White peers [15]. Collectively, these findings show how T2I systems not only reproduce but often intensify existing stereotypes, raising parallel concerns about how representational patterns surface in other generative modalities such as text.

2.3 Representational Bias in Text Generation

Similar representational distortions are found in text generation. Story completions often describe women in relation to appearance or family, while men are framed in terms of power and leadership [47]. In simulated workplace interactions, women and racial minorities receive negotiation advice that encourages them to ask for less and use more deferential language [31]. Nationality bias appears as well, with more negative associations generated for countries with lower global visibility [69]. In news content, LLMs replicate racial and gender disparities, such as producing fewer female-specific terms than found in comparable New York Times or Reuters articles, and more negative sentiment when describing Black individuals [25].

Some of these patterns trace back to the building blocks of LLMs. Word embeddings trained on large text corpora encode gender and occupational stereotypes, producing analogies such as "*man is to computer programmer as woman is to homemaker*" [14]. These associations persist in contemporary LLMs, surfacing in creative writing, political statements, and persona generation. Even when demographic labels are balanced, LLMs may apply group-specific linguistic tropes: for example, consistently describing Latina women as "vibrant" or "curvaceous" [18]. Collectively, this evidence shows that representational bias is not confined to images but extends into text, raising concerns about how occupations and social groups are portrayed at scale. Collectively, this evidence shows that representational bias is not confined to images but extends into text, raising concerns about how occupations and social groups are portrayed at scale.

More recently, benchmarking studies have begun to analyze LLM predictions for occupations directly. The OCCUGENDER benchmark shows that models systematically align male- and female-dominated jobs with stereotypical genders, predicting male pronouns for male-dominated roles and female pronouns for female-dominated roles [17]. This work provides an important foundation for understanding how occupational stereotypes are encoded in LLM outputs, and also highlights opportunities to examine how such representational patterns extend beyond gender, into other demographic dimensions and into more applied contexts such as persona generation.

This work is currently under peer review for publication. This version is a preprint and has not undergone formal peer review.

2.4 Generative Personas as a Site of Representational Bias

Personas were formalized by Alan Cooper in the 1990s as fictional yet evidence-based characters, grounded in interviews and surveys, to help designers move beyond vague notions of "the user" [19]. Since then, personas have been adopted not only in HCI and UX but also in marketing and product development as a way to make user groups tangible and actionable [49, 52]. Critiques, however, note that personas often drift from empirical grounding and instead collapse into reductive stereotypes that reflect designers' assumptions rather than actual users [16].

Generative AI extends this tension in new ways. Like traditional personas, LLM-generated profiles are ostensibly "grounded in data," but here the data consists of massive and opaque training corpora and post-training adjustments rather than explicit interviews or surveys. This means that, in our study, model outputs reflect whatever demographic assumptions are already embedded in the training process, rather than being tied to newly collected user data. Other researchers have pursued hybrid approaches: for example, the PersonaCraft pipeline shows how survey datasets can be transformed into "humanized" personas with high ratings for clarity and credibility [39]. In contrast, our approach evaluates representational patterns that emerge when models are prompted without external grounding—revealing the stereotypes and systematic skews that surface directly from model training.

Recent empirical work suggests both the promise and risks of this shift. LLM-generated personas have been judged by expert raters as broadly comparable in quality to human-crafted personas, though results vary by dimension and model, underscoring the need for oversight and transparency (Smrke et al., 2025). Other experiments show that adding synthetic images to text-based personas does not significantly alter user perceptions, suggesting that textual narratives themselves carry the strongest representational signals [58].

Together, these findings indicate that generative personas can be convincing and even useful, but their reliance on opaque training data raises the possibility of reinforcing stereotypes rather than faithfully representing users. This makes them a critical site for investigating representational bias. In the next section, we outline five models of representation that provide a framework for evaluating how such outputs align with—or diverge from—real-world baselines.

2.5 Models of Representation

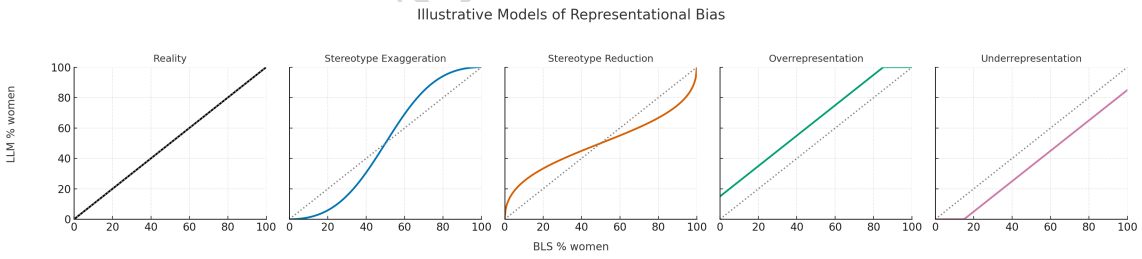

Fig. 2. Conceptual illustration of how different bias patterns can appear when comparing large language model (LLM) outputs to BLS data on gender representation in occupations. Each panel plots BLS % women (x-axis) against LLM % women (y-axis). The dotted gray diagonal indicates perfect parity (LLM = BLS). The five bias patterns are: Reality (accurate representation), Stereotype Exaggeration (amplified extremes, slope > 1 in logit space), Stereotype Reduction (dampened extremes, slope < 1 in logit space), Overrepresentation (uniform upward shift), and Underrepresentation (uniform downward shift). These schematic curves are illustrative rather than empirical, and are shown separately for clarity; in practice, models may exhibit combinations (e.g., overall underrepresentation combined with stereotype exaggeration).

Kay et al. [41] demonstrated that Google Image Search not only underrepresented women in occupational results but also exaggerated existing gender stereotypes, shifting perceptions of who works in those fields. They introduced three models of representation: a *reality model* (outputs mirror real-world data), a *stereotype model* (outputs exaggerate demographic skews), and a *balanced model* (outputs equalize representation across groups). Building on this typology and more recent audits of generative systems [12, 35, 50], we extend the framework to capture five systematic patterns in large-scale persona generation (Fig. 2).

(1) **Reality.** Outputs align closely with external benchmarks such as the U.S. Bureau of Labor Statistics. This preserves accuracy but can also naturalize inequities when occupations are already imbalanced. Thus, while BLS offers a convenient benchmark, it should be understood as a descriptive baseline rather than a normative ideal.
(2) **Stereotype Exaggeration.** Outputs amplify demographic skews, making male- or female-dominated occupations appear even more segregated. Evidence suggests that users sometimes rate stereotypical outputs as higher quality because they match expectations [41], highlighting tensions between perceived realism and equitable representation. Similar amplification has been observed across text-to-image systems and LLM narratives [1, 47].
(3) **Stereotype Reduction.** Outputs dampen extremes, portraying strongly gender- or race-skewed occupations as more balanced. While superficially fair, this creates "surface fairness" without addressing deeper disparities [65].
(4) **Overrepresentation.** Outputs consistently elevate certain groups above their real-world share, independent of occupation. Media audits have documented this for Hispanic workers [50, 64], paralleling "diversity washing" in corporate contexts [6].
(5) **Underrepresentation.** Outputs systematically suppress certain groups across roles, such as the consistent erasure of women and racial minorities in AI-generated professional imagery [32, 43].

These five models provide a structured lens for analyzing how generative personas diverge from or align with demographic baselines. In our study, we operationalize them through regression estimates: intercepts (α) capture systematic over- or underrepresentation, while slopes (β) capture exaggeration or reduction of demographic skews. This framework positions us to test not only whether biases exist but also how they systematically take shape across occupations and models.

Beyond differences across occupations, an open question is whether representational patterns diverge across models themselves. Algorithms do not operate in a cultural vacuum: research shows that they inevitably embed the institutional and societal values of their creators and regulators [3, 10, 29]. Because large language models differ in organizational origin, training data, and post-training safety objectives, we might expect provenance to leave a measurable imprint on representational outcomes [38, 48]. For instance, DeepSeek has refused to answer politically sensitive questions about Tiananmen Square, reflecting constraints in the Chinese context [70], whereas U.S.-based models may be more attentive to issues of race and gender identity that are particularly salient in American public discourse [4]. These considerations set the stage for our cross-model analysis, where we compare bias patterns across LLMs trained under different institutional and cultural regimes.

3 Methods

3.1 Research Design and Framework

We conducted a large-scale audit of occupational personas generated by four widely used large language models (LLMs). Our design adapts approaches developed in algorithmic audit studies that compare media outputs to real-world demographic baselines [41, 50]. In these studies, regression analysis of representation against BLS data provided evidence

of systematic over- or underrepresentation and stereotype exaggeration. Following this precedent, we applied similar techniques to LLM-generated occupational personas to evaluate both average bias and systematic distortions.

3.2 Occupation and Benchmark Data

The benchmark dataset was drawn from the U.S. Bureau of Labor Statistics' Current Population Survey Annual Averages for 2023 [68]. We analyzed 41 occupations, selected to align with prior audit studies [41] while ensuring sufficient sample sizes (at least 50,000 employed persons). Several occupations were re-named or substituted with equivalent 2023 categories where prior categories were retired.

To allow direct comparison with BLS data, we restricted analysis to binary gender (male, female) and four racial/ethnic groups (White, Black, Asian, Hispanic). For race, models could assign one or more categories, enabling multi-racial identities. Intersectional (e.g., Black women) and non-binary identities could not be benchmarked because such categories are not reported in CPSAA.

3.3 Persona Generation Procedure

Each model was prompted with the instruction: "Generate a profile for: <CAREER TERM> in the United States." Outputs were returned in JSON with seven fields (name, age, gender, ethnicity, salary, motivations, biography). No balancing instructions were provided, allowing models to rely solely on their internal training distributions. The full prompt text and JSON schema are provided in the Supplementary Materials.

For each occupation, we generated 10,000 profiles per model. DeepSeek was limited to 1,000 profiles due to API constraints, as its interface does not support efficient batching of requests. While this yields fewer samples for DeepSeek relative to other models, the volume is sufficient to detect systematic patterns, and our regression-based analyses weight occupations rather than raw sample size. Requests were issued in independent context windows to prevent models from balancing outputs across prompts. Default API temperature settings were used[1]. Less than 1% of responses were malformed (e.g., missing fields or incorrect data types); these were discarded and regenerated to maintain equal sample sizes across occupations.

We used a single, fixed prompt template to maintain comparability across occupations and models. Prior work (e.g., "OCCUGENDER" [17]) has shown that prompt wording can substantially affect representational outcomes. Our approach therefore establishes a controlled baseline, while acknowledging that findings are bounded to this elicitation regime. We return to the limitations of the single-prompt design, and the value of multi-prompt auditing strategies, in the Limitations section.

Table 1. Profile of LLMs Selected for Persona Generation

LLM Version	Organization	Country of Origin	U.S. Market Share (August 2025) [26]	AI Safety Grade (Summer 2025) [28]
GPT 4	OpenAI	USA	60.40%	C
Gemini 2.5	Google	USA	13.50%	C
Deepseek V3.1	DeepSeek	China	0.4%	F
Mistral medium	Mistral AI	France	Not listed	Not listed

[1]We use default settings to approximate typical end-user or API integrator behavior rather than to endorse these settings as optimal for fairness.

3.4 Models Audited

We evaluated four LLMs that differ in organizational origin and stated safety commitments: GPT-4 (OpenAI, USA), Gemini 2.5 (Google, USA), DeepSeek V3.1 (DeepSeek, China), and Mistral Medium (Mistral AI, France). As discussed in Section 2.5, provenance may shape representational outcomes because models embed institutional and cultural values through their training data and post-training objectives. Table 1 summarizes the versions, organizational origins, U.S. market shares, and safety grades of the models included in our audit.

Market share figures are from First Page Sage's report on generative AI chatbots [26]. Safety scores are drawn from the AI Safety Index (Summer 2025), published by the Future of Life Institute with input from a panel of independent experts [28]. The Index aggregates 33 indicators across six domains, combining results from external benchmarks (e.g., HELM, TrustLLM, UK AISI red-teaming) with company documentation and disclosures. Grades follow a U.S.-style GPA system (A–F) intended to make relative differences accessible to a broad audience, though they reflect primarily U.S.-based perspectives on responsible AI practices.

Our focus on four models reflects practical constraints of time and computational cost. While a broader set of models would provide greater coverage, these four represent a meaningful spread in provenance: two U.S.-based firms with large market share, alongside one Chinese and one European entrant with smaller presence. Together, they allow us to assess both convergence and divergence in representational patterns across models developed under different institutional and cultural regimes.

3.5 Statistical Analysis

We compared LLM outputs against BLS baselines using two complementary approaches.

Average bias. We calculated percentage-point differences between model outputs and BLS distributions for each occupation and demographic group. Negative values indicate underrepresentation; positive values indicate overrepresentation.

Regression framework and robustness. To assess systematic patterns, we estimated generalized linear logistic regressions of model-generated proportions on BLS proportions. Following prior audit studies of occupational bias [41, 50], we interpret the regression intercept (α) as systematic over- or underrepresentation and the slope (β) as exaggeration or reduction of demographic skews (amplification when slope > 1, dampening when slope < 1).

We chose to center regressions at each group's median BLS share so that α reflects bias where the bulk of the data lies. This avoids extrapolation into occupations with unusually high or low demographic shares and improves interpretability of regression estimates.

To assess robustness, we estimated three specifications: (1) raw regression, (2) trimmed regression excluding the top and bottom 5% of occupations by BLS representation, and (3) robust regression, which downweights high-leverage occupations. Robust regression, widely used in social science to minimize distortion from outliers [34, 45], is reported as our primary specification. Visual regression lines are presented in probability space, while numerical estimates of α and β are reported on the centered logit scale.

Multiple comparisons. To control for inflated false positives, we applied false discovery rate (FDR) correction across regression tests [9]. Significance levels are reported with stars ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$), and symbol codes (e.g., +++, −) summarize direction and magnitude.

This work is currently under peer review for publication. This version is a preprint and has not undergone formal peer review.

Interpretation. We interpret regression intercepts (α) as evidence of systematic over- or underrepresentation and regression slopes (β) as evidence of stereotype exaggeration or reduction. To make these results more interpretable, we translated raw logit coefficients into percentage-point shifts (for α) and slope deviations relative to parity (for β), then assigned symbol codes to indicate the magnitude and direction of effects.

For systematic shift (α), we expressed results as percentage-point differences from BLS at the group's median occupational share. Because effects in our data ranged from negligible (<0.5 pp) to very large (>10 pp), we classified them as: nonsignificant (0), small (+/−, 0.5–3 pp), moderate (++/−, 3–10 pp), and strong (+++/—, >10 pp). This scaling allows us to highlight both modest distortions and extreme systematic shifts.

For stereotype slope (β), we measured the deviation of regression slopes from parity (1.0). Prior work (e.g., [41, 47]) has typically emphasized the presence or absence of stereotype exaggeration, often with small thresholds. In our data, however, observed slope deviations ranged from near zero to >2.5. To capture this wider spread while retaining comparability with prior work, we classified slope deviations as: nonsignificant (0), small (+/−, 0.1–0.5), moderate (++/−, 0.5–1.5), and strong (+++/—, >1.5).

Together, these coding rules provide a consistent way to summarize regression estimates across dozens of group × model comparisons, while distinguishing mild exaggerations from extreme distortions.

3.6 Transparency and Reproducibility

In total, our dataset comprises over 1.5 million generated profiles across four models and 41 occupations. We provide full regression tables, the exact prompt text, the JSON field schema, and parsing scripts in the supplementary materials. All synthetic data analyzed in this study was generated by LLMs, with no additional post-processing beyond formatting and error correction.

3.7 Ethical Considerations

This study did not involve human participants and therefore did not require ethical approval. All data analyzed either consisted of synthetic outputs generated by AI models or publicly available aggregate statistics. Nonetheless, we recognize that representational biases in AI outputs can perpetuate harms, particularly for minority groups. Our analyses are designed to surface such biases, and we aim to approach their interpretation with care and sensitivity.

4 Results

4.1 RQ1: Over- and Underrepresentation by Occupation

We first examine whether AI-generated personas systematically over- or underrepresent demographic groups across specific occupations. This analysis focuses on average deviations from U.S. labor market distributions, identifying where models most strongly distort gender and racial representation.

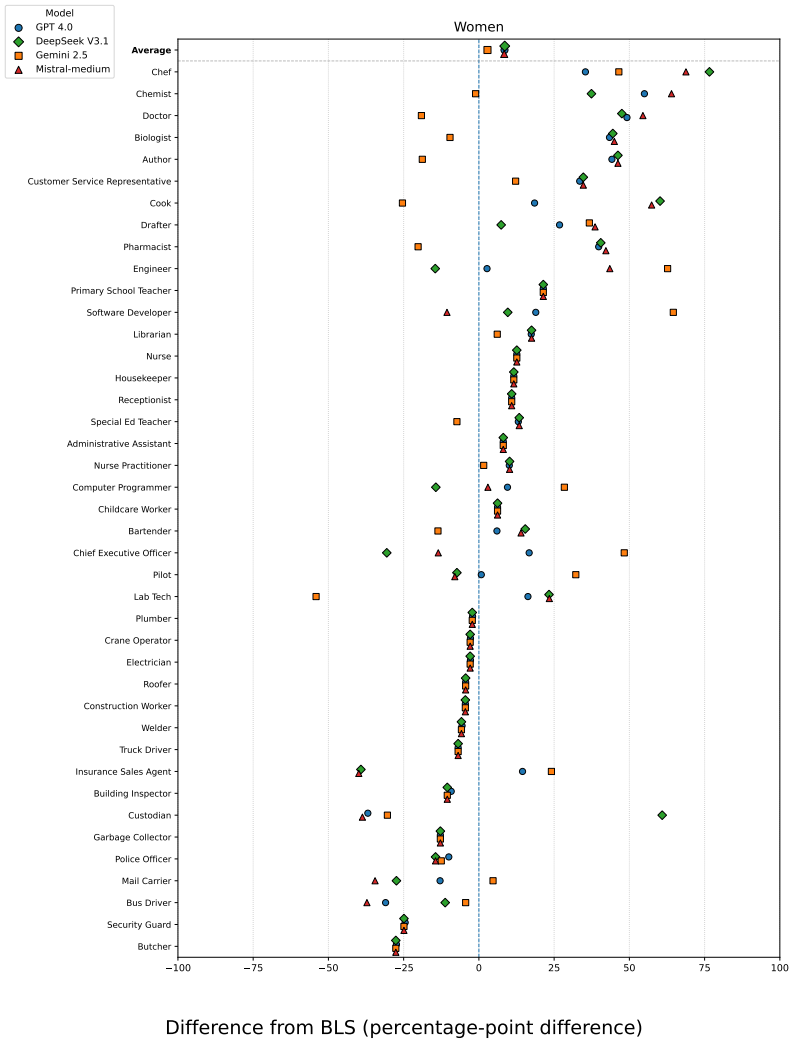

Fig. 3. Differences between four large language models' representations of gender in occupational prompts compared to BLS benchmarks. Occupations are shown on the y-axis, with an "Average" row at the top summarizing mean differences across all 41 occupations. The x-axis shows percentage-point (pp) differences from BLS (absolute, not relative): negative values indicate underrepresentation and positive values indicate overrepresentation, with the dashed vertical line marking parity (0). Occupations are ordered from those where women are most overrepresented to most underrepresented. Only women are shown, with men serving as the implicit complement. Each dot represents one occupation, and points that would otherwise overlap are jittered vertically for readability.

Gender. Across occupations, women are *modestly overrepresented on average*, though patterns vary widely (Figure 3). Overrepresentation is strongest in high-status professional roles associated with intellectual rigor and socioeconomic

success (e.g., *Engineer*, *Software Developer*, *Author*, *Biologist*). Women are also overrepresented in traditionally female-majority occupations (e.g., *Nurse*, *Childcare Worker*, *Primary School Teacher*, *Housekeeper*, *Cook*), though by smaller margins, consistent with their already high BLS baselines. Conversely, men are most overrepresented in blue-collar occupations aligned with traditional masculine stereotypes (e.g., *Truck Driver*, *Construction Worker*, *Electrician*). Taken together, these patterns show that LLMs amplify women's presence in both professional and stereotypically feminine roles, while consolidating men's association with manual labor.

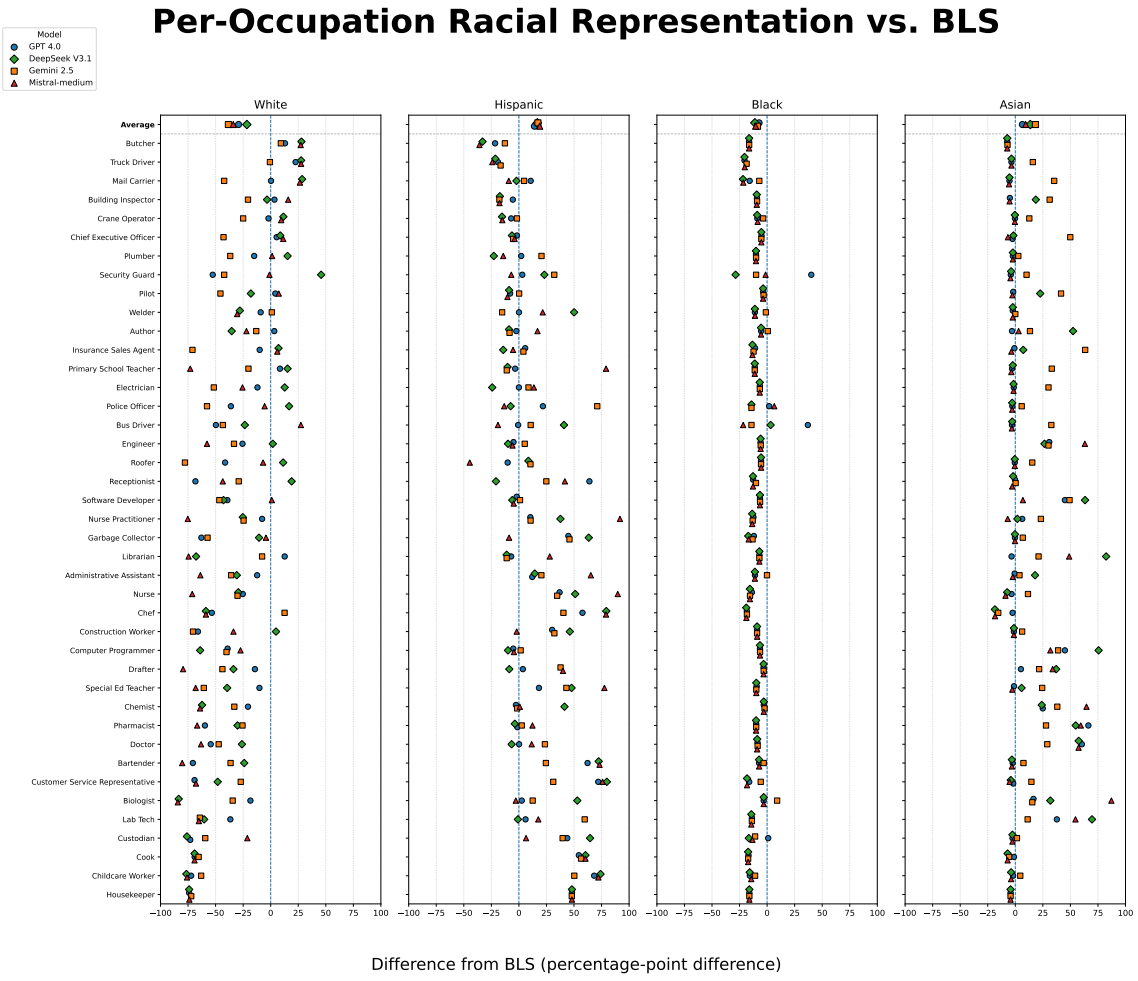

Fig. 4. Differences between four large language models' representations of racial groups in occupational prompts compared to BLS benchmarks. Occupations are shown on the y-axis, with an "Average" row at the top summarizing mean differences across all 41 occupations. The x-axis shows percentage-point differences from BLS (absolute, not relative): negative values indicate underrepresentation and positive values indicate overrepresentation, with the dashed vertical line marking parity (0). Occupations are ordered from those where White workers are most overrepresented to those where they are most underrepresented. Separate panels show results for White, Hispanic, Black, and Asian workers. Each dot represents one model (ChatGPT = circle, Gemini = square, DeepSeek = diamond, Mistral = triangle); points that would otherwise overlap are jittered vertically for readability.

Race. Racial representation shows sharper and more systematic distortions (Figure 4). On average across occupations and across all four models, White workers are underrepresented by roughly −31 percentage points, Black workers by −9 points, while Hispanic (+17 points) and Asian (+12 points) workers are systematically overrepresented. Importantly, the implications of these shifts differ given group baselines in the U.S. workforce. Although White workers show the largest raw distortion in pp terms, they remain the numerical majority in most occupations. By contrast, a −9 point shift for Black workers represents a far larger share of their baseline presence, nearly erasing them from many occupations except a handful of stereotyped roles (e.g., *Security Guard* and *Bus Driver* in GPT-4).

At the occupation level, distortions are often extreme:

- **Hispanic Housekeepers (all models).** Across GPT-4, Gemini, DeepSeek, and Mistral, Hispanic workers are inflated by about +48 points relative to BLS. This is a cross-model stereotype of Hispanics in low-status service work.
- **Hispanic overrepresentation in care and service.** *Nurse Practitioner* (+92 points, Mistral), *Nurse* (+90 points, Mistral), *Customer Service Representative* (+80 points, DeepSeek), and *Chef* (+79 points, DeepSeek) all display extreme Hispanic inflation.
- **Asian inflation in scientific/professional roles.** *Biologist* (+87 points, Mistral) and *Librarian* (+82 points, DeepSeek) highlight strong overrepresentation.
- **White underrepresentation.** White workers are sharply suppressed in *Biologist* (−84 points, Mistral; −83 points, DeepSeek), *Bartender* (−80 points, Mistral), and *Drafter* (−79 points, Mistral).

Taken together, RQ1 shows that LLMs do not merely mirror occupational demographics but impose *directional, group-specific distortions*. These include amplifying women in professional and feminine-coded roles, underrepresenting Black workers to near-erasure, systematically inflating Hispanic and Asian representation, and sharply suppressing White presence in select occupations. The consistency of these distortions across four different models suggests they are embedded in shared training data and conventions, rather than model-specific quirks.

4.2 RQ2: Systematic Deviations and Stereotype Exaggeration

We next examine whether AI-generated personas systematically deviate from labor market distributions, focusing on stereotype exaggeration — cases where models amplify existing demographic skews across occupations.

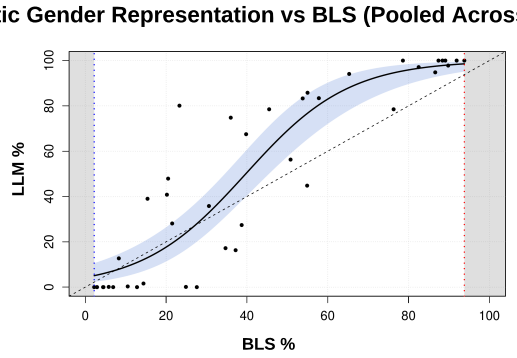

Fig. 5. Systematic gender representation pooled across all models. Each point is an occupation, plotting BLS percentage of women against pooled LLM output. The diagonal line indicates parity. The fitted curve follows an S-shaped pattern, showing that models tend to exaggerate gender skews at both male- and female-dominated ends of the distribution.

Table 2. Systematic Gender Representation (Pooled). Logistic regression estimates of systematic bias for gender representation pooled across 41 occupations. Over/underrepresentation is based on the intercept (α), and stereotyping is based on the slope (β). Stars indicate FDR-adjusted significance ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$). For significant results, symbol codes (≈0, +, ++, +++, −, −−, −−−) summarize the direction and strength of effects.

	Systematic Shift			Stereotype Slope		
Group	**α (logit)**	**Systematic shift (pp)**	**Interpretation**	**β (logit)**	**Slope deviation**	**Interpretation**
Women	−0.85	−6.14***	− −	13.26	+0.17***	+

Gender. Figure 5 plots the share of women generated by the models for each occupation against BLS baselines, pooling results across all models. The diagonal line marks perfect parity: values on the line indicate accurate representation, values below underrepresentation, and values above overrepresentation. The fitted curve follows an S-shaped pattern: at the left end of the distribution, corresponding to male-dominated occupations, the models consistently underpredict women; through the middle and higher range of BLS values, the curve rises above parity, indicating overprediction of women; and at the far right it flattens at the ceiling, further exaggerating female dominance. In short, the models amplify gender imbalances, producing too few women in male-dominated fields and too many in female-dominated fields.

Table 2 quantifies these patterns using robust regression. At the median occupation (about 36% women in BLS data), models underrepresent women overall (−6.1 percentage points, FDR-significant). The slope deviation is positive (+0.17), indicating stereotype exaggeration: the gap between models and BLS grows as occupations become more strongly segregated. Taken together, the regression highlights a central tendency toward underrepresentation at the median, while the S-shaped curve shows how this systematic shift coexists with local overrepresentation in mid-to-high BLS occupations. Both perspectives converge on the same point: LLMs exaggerate gender segregation rather than reducing it.

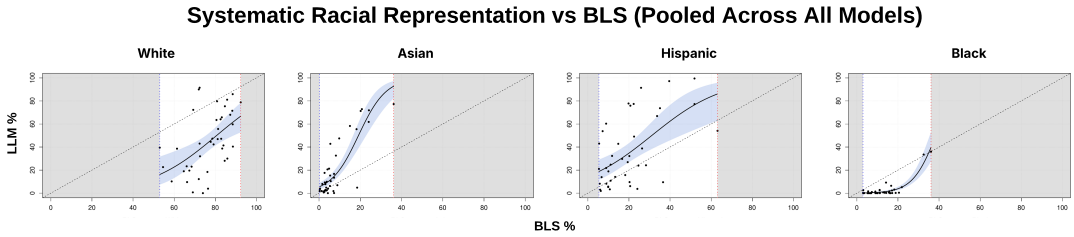

Fig. 6. Systematic racial representation pooled across all models. Each point is an occupation, plotting BLS percentages of racial groups against pooled LLM outputs. The diagonal line indicates parity. The fitted curves show underrepresentation of Black and White workers, overrepresentation of Hispanic and Asian workers, and stereotype exaggeration of demographic skews.

Table 3. Systematic Racial Representation (Pooled). Logistic regression estimates of systematic bias for racial representation across 41 occupations. Notation as in Table 2.

	Systematic Shift			Stereotype Slope		
Group	**α (logit)**	**Systematic shift (pp)**	**Interpretation**	**β (logit)**	**Slope deviation**	**Interpretation**
Asian	−2.47	3.37^{***}	+ +	18.89	$+2.10^{***}$	+ + +
Black	−5.81	-10.80^{***}	− − −	21.21	-0.89^{***}	− −
Hispanic	−0.96	9.40^{***}	+ +	6.93	$+0.43^{***}$	+
White	−0.14	-32.10^{***}	− − −	9.65	$+1.13^{***}$	+ +

Race. Figure 6 and Table 3 extend this analysis to racial categories. Here, the distortions vary across groups. Black workers are consistently underrepresented: the regression shows a large negative shift (−10.8 pp) and a slope below parity (−0.89), indicating stereotype reduction through near-erasure at typical occupational levels, with presence surfacing only in a handful of outlier occupations (such as *Security Guard* and *Bus Driver*).

White workers are also underrepresented overall (−32 pp), but with a positive slope (+1.13), meaning that their share is exaggerated in occupations where they are already more common. By contrast, Hispanic (+9.4 pp) and Asian workers (+3.4 pp) are systematically overrepresented. For both groups, positive slope deviations (Hispanic +0.43; Asian +2.10) indicate stereotype exaggeration, with the gap between models and BLS widening in occupations where these groups already have higher representation.

Together, these findings show that models do not simply mirror real-world proportions. Instead, they impose systematic distortions: erasing some groups in the middle of the distribution, while amplifying overrepresentation or exaggeration for others.

4.3 RQ3: Cross-Model Comparisons of Systematic Bias

We then compare how these representational biases differ across models. This analysis asks whether some models show stronger exaggeration, reduction, or consistent shifts than others, and whether these differences align with organizational origins or safety commitments.

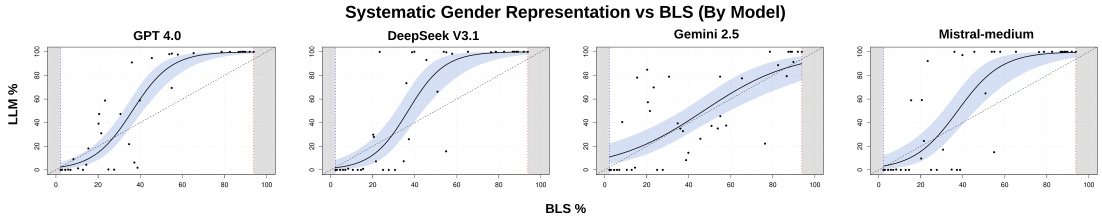

Fig. 7. Systematic gender representation by model. Each panel plots BLS percentages of women in occupations against LLM outputs, with parity shown as the diagonal line. All models exhibit an S-shaped pattern, indicating stereotype exaggeration. GPT-4, DeepSeek, and Mistral show modest underrepresentation of women overall, while Gemini is closer to parity but still exaggerates gender skews at the extremes.

Table 4. Systematic Gender Representation (By Model). Logistic regression estimates of systematic bias for gender representation, disaggregated by model. Notation as in Table 2.

		Systematic Shift			Stereotype Slope		
Model	**Group**	α **(logit)**	**Systematic shift (pp)**	**Interpretation**	β **(logit)**	**Slope deviation**	**Interpretation**
ChatGPT	Women	−0.13	10.72***	+ + +	14.91	+0.17***	+
DeepSeek		−0.22	9.12***	+ +	27.61	+0.19***	+
Gemini		−0.93	−7.62***	− −	6.81	+0.05***	≈0
Mistral		−0.29	6.92***	+ +	19.41	+0.18***	+

Gender. Figure 7 shows a clear split. Three providers—ChatGPT, Mistral, and DeepSeek—*lift women on average* (+10.7 pp, +6.9 pp, and +9.1 pp, respectively), while *Gemini* is the only model that *lowers women* around the median point (−7.6 pp). Across all four, slopes are small, indicating little additional steepening or flattening of gender segregation by occupation. In practice, readers see two patterns: a general "lift" for women in three models with minimal stereotyping, and an "offset down" in Gemini, also with minimal stereotyping.

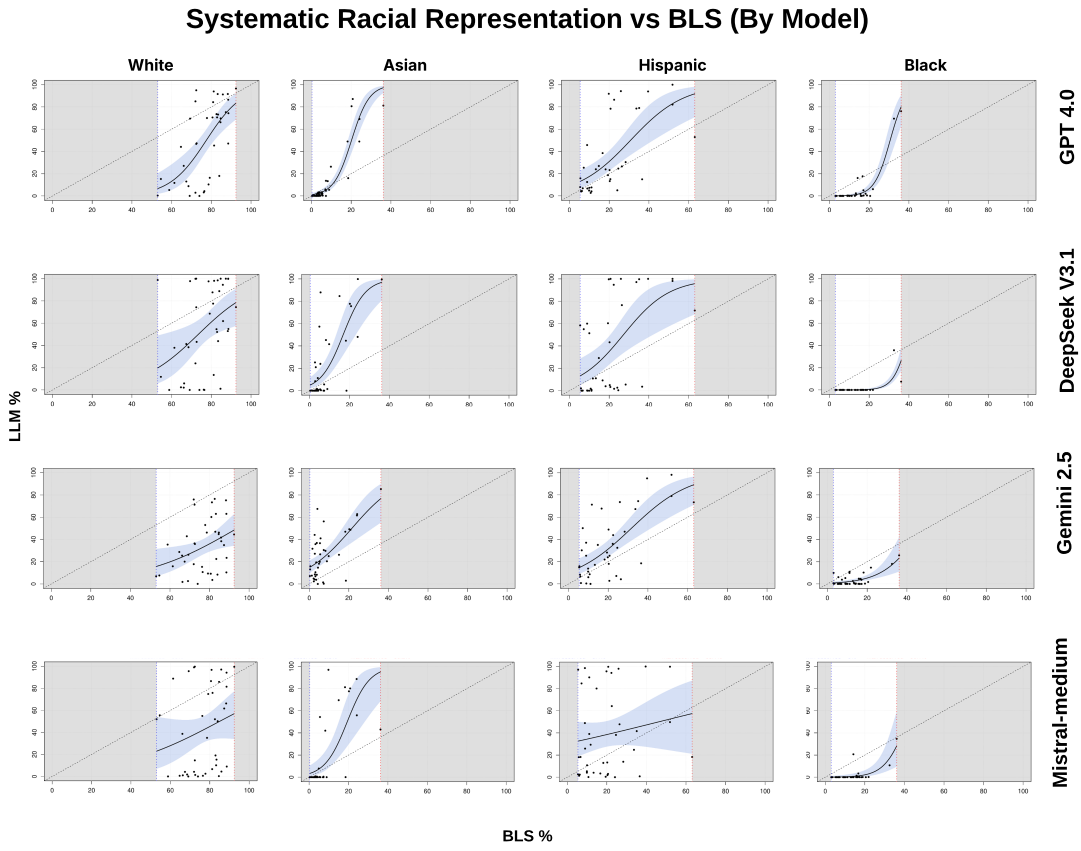

Fig. 8. Systematic racial representation by model. Each panel plots BLS percentages of racial groups against LLM outputs, with parity shown as the diagonal line. All models display clear bias patterns: Black workers are consistently underrepresented, Hispanic workers are overrepresented, and Asian and White workers show stereotype exaggeration. Gemini is slightly closer to parity than GPT-4, DeepSeek, or Mistral, but still departs from BLS distributions.

Race/Ethnicity. Figure 8 highlights both broad consistencies and meaningful differences across groups.

- **White.** All models underrepresent White workers on average, but the magnitude varies widely: Gemini shows the largest decline (−44.3 pp), Mistral is also large (−34.6 pp), ChatGPT is moderate (−23.7 pp), and DeepSeek the smallest (−17.7 pp). Slope behavior diverges: ChatGPT and DeepSeek concentrate White representation where it is already high (positive slopes), whereas Gemini is near flat (≈ 0), removing White presence more evenly across occupations.
- **Asian.** Only Gemini overrepresents on average (+12.6 pp). ChatGPT, Mistral, and DeepSeek slightly underrepresent (−2.8 to −4.1 pp), yet all four models concentrate Asian representation in occupations where Asians are already more common (positive slopes), with the strongest concentration in DeepSeek and Mistral. The net effect is a contrast between Gemini's broader uplift and the other providers' sharper clustering.

Table 5. Systematic Racial Representation (By Model). Logistic regression estimates of systematic bias for racial representation, disaggregated by model. Notation as in Table 2.

		Systematic Shift			Stereotype Slope		
Model	**Group**	α **(logit)**	**Systematic shift (pp)**	**Interpretation**	β **(logit)**	**Slope deviation**	**Interpretation**
ChatGPT	Asian	−4.14	-2.83^{***}	−	27.98	$+2.03^{***}$	+ + +
	Black	−6.45	-10.94^{***}	− − −	33.07	-0.85^{***}	− −
	Hispanic	−1.21	4.68^{***}	+ +	10.24	$+0.91^{***}$	+ +
	White	0.20	-23.71^{***}	− − −	11.65	$+1.52^{***}$	+ + +
DeepSeek	Asian	−5.04	-3.71^{***}	− −	37.43	$+2.80^{***}$	+ + +
	Black	−10.91	-10.60^{***}	− − −	33.23	-1.00^{***}	− −
	Hispanic	−1.49	-0.46^{***}	≈0	27.65	$+2.16^{***}$	+ + +
	White	0.45	-17.72^{***}	− − −	10.13	$+1.24^{***}$	+ +
Gemini	Asian	−1.58	12.62^{***}	+ + +	10.41	$+1.05^{***}$	+ +
	Black	−5.71	-10.77^{***}	− − −	18.70	-0.90^{***}	− −
	Hispanic	−1.10	6.79^{***}	+ +	7.92	$+0.56^{***}$	+ +
	White	−0.65	-44.26^{***}	− − −	4.35	-0.06^{***}	≈0
Mistral	Asian	−5.89	-4.12^{***}	− −	42.07	$+2.67^{***}$	+ + +
	Black	−10.81	-11.10^{***}	− − −	40.70	-1.00^{***}	− −
	Hispanic	−0.81	12.48^{***}	+ + +	2.45	-0.47^{***}	−
	White	−0.24	-34.64^{***}	− − −	5.83	$+0.36^{***}$	+

- **Hispanic.** Models split on both average level and concentration. Mistral shows the largest uplift (+12.5 pp) and reduces stereotyping (negative slope), distributing Hispanic personas more broadly. Gemini (+6.8 pp) and ChatGPT (+4.7 pp) also raise average presence and show moderate concentration (positive slopes). DeepSeek is near zero on average (−0.5 pp) yet displays the strongest concentration (largest positive slope), clustering Hispanic personas where BLS shares are already high.
- **Black.** The pattern is uniform: each model underrepresents by about −11 pp (ChatGPT −10.9; Gemini −10.8; Mistral −11.1; DeepSeek −10.6) and exhibits stereotype reduction (negative slopes around −0.85 to −1.00). Substantively, this reflects near-erasure across the typical BLS range: representation remains low even where Black shares are higher, with visibility mainly in a few outlier occupations.

Overall takeaways. Across providers, there is a *shared backbone*: Black and White workers are underrepresented on average, gender slopes are small, and Asian representation is concentrated where it is already high. Within that common pattern, however, models diverge in both the midpoint (who is present or absent on average) and the degree of concentration (how sharply representation clusters by occupation). Most notably, three models lift women while Gemini lowers them; White underrepresentation ranges from moderate to extreme with different slope behavior; Mistral broadly uplifts Hispanics while reducing stereotyping, whereas DeepSeek shows near-zero average change but strong concentration; and Gemini uniquely boosts Asian averages while others underrepresent but cluster them. Importantly, these differences do not align cleanly with country of origin or stated safety posture: the two US models (ChatGPT and Gemini) diverge substantially from each other, and the non-US models (Mistral, DeepSeek) differ as well. In short, provider choice changes who is visible and where, even as the broad directions of distortion remain similar.

5 Discussion

5.1 Interpreting systematic bias

Representational systems do not simply mirror reality; they shape who is seen as belonging in particular roles. HCI research has long shown that bias is encoded into technical systems rather than a neutral byproduct [27], and perspectives such as Critical Race Theory for HCI highlight how representational choices are bound up with power and belonging [53]. Personas exemplify this duality: designed to move beyond vague notions of "the user," they often collapse into stereotypes. Generative personas extend this tension, drawing on opaque training data rather than empirical user studies. Our study builds on Baumer's call for human-centered algorithm design [8], showing how large-scale persona generation encodes systematic demographic assumptions. We analyze these patterns through a five-model framework (reality, exaggeration, reduction, overrepresentation, underrepresentation) that operationalizes systematic shifts (α) and slope patterns (β) against external baselines.

Across occupations, models impose structured distortions rather than simply reproducing distributions. On average, women, Asian, and Hispanic workers are lifted, while White and Black workers are reduced—but the same shift has different implications depending on group baselines. Slope patterns show the hallmark S-curve for gender (fewer women where women are rare, more where they are common) and race-specific amplification or suppression. Apparent "stereotype reduction" for Black workers reflects near-erasure across the typical occupational range rather than equitable balancing. Parallel work finds similar flattening or erasure of marginalized identities, especially non-binary and transgender people [60, 61], though our reliance on BLS data prevented analyzing these groups. Large-scale audits of online images likewise demonstrate that algorithmic representations exaggerate gender gaps [35]. Our findings therefore suggest that LLM persona generation reproduces these same exaggeration and erasure dynamics, but in ways that could potentially become embedded in design and product development workflows, as well as in educational settings (e.g., classroom assignments or training modules), marketing research (e.g., consumer segmentation studies), and creative applications (e.g., worldbuilding and character generation).

5.2 Undesirable and Ambiguous Representativeness

A key finding from RQ1 is that distortions are not only about over- or under-representation, but whether such representations are socially desirable. Some patterns reproduce harmful stereotypes. For example, Hispanic workers are heavily overrepresented as housekeepers by all four models, with GPT-4 generating nearly 100% of profiles as Hispanic and female. Hispanic presence is also inflated in other care and service roles, such as nurse practitioner, chef, and customer service representative. These distortions echo cultural tropes that reduce Hispanic identity to low-status labor [21], with real-world implications for labor market opportunities [72]. This aligns with broader HCI critiques that algorithmic systems often reproduce historical inequities embedded in categories of work and identity [22].

Similar dynamics appear for gender and race. Women are frequently exaggerated in caregiving and domestic roles, reinforcing assumptions about what kinds of work they "ought" to do, despite decades of feminist design work challenging such essentialist framings [22]. Black workers are generally underrepresented, but when present, are clustered in roles such as *Security Guard* or *Bus Driver*, echoing historic racial tropes.

At the same time, not all deviations from labor market baselines are harmful. In our data, women are often overrepresented in professions associated with intellectual rigor and socioeconomic success (e.g., biologist, author, computer programmer). Given persistent gender gaps in STEM and leadership, such inflation could be read as aspirational. Similarly, increased representation of Hispanic or Asian workers in roles that diverge from historic stereotypes may broaden

imagined boundaries of belonging. Conversely, White workers, still the numerical majority of the U.S. workforce, are consistently underrepresented. While this departs from demographic fidelity, it could be interpreted as rebalancing visibility toward marginalized groups.

Together, these findings highlight the **ambiguity of representativeness** in generative AI outputs. Distortions can be harmful, neutral, or socially beneficial depending on context. Our study treats all deviations from BLS baselines as systematic distortions, which makes patterns detectable, but it cannot answer when deviations should be celebrated, mitigated, or ignored. This interpretive challenge echoes prior HCI work on search and image audits that showed how representational distortions influence user perception and reinforce stereotypes [50], underscoring the ongoing need for human-centered approaches to algorithm design [8, 42]. Because the normative meaning of deviation is often context-dependent, downstream tools should make representational choices both visible and adjustable. We return to this in § 5.4, where we outline UI mechanisms that support such visibility and adjustment in practice.

5.3 Provider choice has material, distributional consequences

A central takeaway from RQ3 is that providers share a common backbone of distortion (for example, underrepresentation of Black and White workers; concentrated Asian representation) but diverge in offsets and concentrations (such as women's mid-range presence or Hispanic uplift patterns). Crucially, these differences do not align cleanly with country of origin or with models' reported safety scores or reputational claims. For HCI readers, the implication is that provider choice materially changes representational outcomes; "safety" branding alone is not predictive of direction or magnitude of bias. This underscores the need for model-specific audits before any downstream design or policy decision. Industry practitioners themselves have called for concrete fairness auditing tools rather than assurances [36]. Public audits have also been shown to drive vendor behavior change, as in commercial face recognition systems [57]. Yet evaluation gaps remain widespread [37], and recent work argues for disaggregated reporting across subgroups as a baseline for accountability [7]. Our cross-model analysis contributes to this line of work by demonstrating how regression-based audits can surface systematic shifts (α) and slope distortions (β) across providers, making visible where model choice alters who is represented.

5.4 From diagnosis to design: UI mechanisms for governing representation

If audits show where representational patterns diverge, the next step is to consider how these insights might shape the tools themselves. Our results highlight patterned distortions across occupations that recur across providers, suggesting that persona tools could treat representation as a first-class design parameter rather than a hidden byproduct. Prior HCI guidelines emphasize that human–AI interfaces can expose system behavior, support appropriate reliance, and give users ways to recover from errors [2, 73]. Building on this, we outline several UI mechanisms that could be implemented in persona-generation tools used in design research and product development, in educational or creative platforms, in enterprise dashboards for marketing or auditing, and in simulation or world-building contexts where hundreds of personas may be generated at scale [55]. In each setting, the goal is to make representational patterns visible and governable rather than hidden.

Baseline visibility at a glance. Each generated set could include compact meters that quantify distance from an external baseline (for example, BLS) at the group's median share, with brief tooltips explaining the benchmark. This would anchor interpretation in observable outputs and avoid overpromising "transparency" through unverifiable

rationales. Such design would function like a mini "model card" embedded in the interface, surfacing the assumptions and benchmarks applied [30, 51].

Sets, not single answers. One approach could be to default to a small slate (for example, 4–8 personas) with a coverage bar highlighting who is present or absent, plus alerts if any group falls below a configurable floor. Another approach could be to display a single main persona by default, but include two or three *alternative personas below the fold* in compact cards. This keeps the primary view simple while still surfacing diversity and offering quick access to counter-stereotypical variants. Both approaches align with HCI guidance to provide diverse outputs and allow recovery from errors [2].

Audit views and provider comparison. To sustain visibility over time, tools could provide a batch audit view that summarizes distributions by team and timeframe, compares them to baseline, and surfaces α/β summaries. Prior work shows that such audit dashboards can be valuable to practitioners [36], including in adjacent domains like news sourcing, where the DIANES toolkit demonstrates how dashboards can track diversity of quoted voices in real time [62]. These kinds of views may even drive vendor behavior change [57]. Interfaces could also support provider comparison: a one-click view rendering side-by-side distributions and α/β deltas, so teams can see, rather than assume, how provider choice shifts representation. This connects directly to our RQ3 finding that reported "safety" scores or reputations are not predictive of bias magnitude or direction.

5.5 Transparency as process trace

Output-level summaries alone (e.g., who appears in a persona set) can hide how representational choices were produced. A complementary approach is *process-trace transparency*: making visible the key steps behind any persona, including (i) which external sources were consulted, (ii) how group distributions were set, (iii) how sampling and randomness were applied, and (iv) what plausible alternatives were available. This kind of disclosure functions like a lightweight, embedded model card—focused on inputs and transformations that matter for representation [30, 51].

To keep such transparency usable, disclosures can be *layered*: a one-line summary on the surface (sources, targets, timestamp) with expandable details on demand. Layering supports appropriate reliance without overload, consistent with HCI guidelines for human–AI interaction [2] and empirical findings that too much detail can overwhelm practitioners or lead to misplaced trust [40, 56]. The form of the disclosure also matters: narrative rationales can make model behavior more understandable to non-experts [23], especially when paired with compact visual cues and numerical summaries.

Design Vignette: Grok 4 and Process-Trace Transparency. In exploratory testing, we prompted Grok 4 (another LLM, not part of our main evaluation) to generate a persona.[2] Rather than outputting a single result, Grok (1) searched for demographic baselines on external sites, (2) generated a Python script to randomize across gender and racial distributions, and (3) surfaced this process to the user, including probability weights and the sampling code.

This trace made Grok's assumptions highly inspectable and offered a glimpse of how process-trace transparency might look in practice. At the same time, the generated personas still reflected stereotypes through narrow and culturally marked name lists (e.g., "Tyrone Davis" or "Jamal Washington" for Black men; "Wei Chen" or "Hiroshi Tanaka" for Asian men). This illustrates that while transparency can reveal how outputs are constructed, it does not by itself address the biases embedded in those constructions.

[2]Link to Grok persona generation results: https://grok.com/share/bGVnYWN5_be77d88c-88be-4766-b3e4-1efe89e5febf

Process-trace transparency can also reduce the need for users to form "folk theories" in the absence of cues about system behavior [24]. For HCI, the lesson is that transparency should be structured as layered disclosure—concise summaries up front, expandable details on demand—and paired with scaffolds for interpretation, such as contextual warnings, alternative examples, or editable components. This combination helps users not only see how outputs were generated, but also critically engage with the representational assumptions embedded in them.

5.6 Limitations

Our analysis has several important limitations. First, we benchmarked model outputs against U.S. Bureau of Labor Statistics (BLS) data. This offers a consistent and widely used point of comparison, but it is not a normative standard. BLS distributions themselves reflect historical inequities in the U.S. labor market (e.g., [11, 13]), so parity with BLS does not necessarily equate to fairness, and deviations from BLS do not automatically signal harm. Accordingly, following prior work that highlights how descriptive accuracy can naturalize existing inequities [41, 50], we treat our results as descriptive evidence rather than prescriptive judgments about fairness.

Second, we constrained analysis to the categories reported by BLS: binary gender (male, female) and four race/ethnicity groups (White, Black, Asian, Hispanic). This design choice enables comparability but excludes non-binary identities, finer racial/ethnic distinctions (e.g., Pacific Islander, Native American, or subgroups within Asian or Hispanic populations), and mixed-race individuals, including populations that may face particular forms of marginalization. These omissions limit what we can conclude about representational harms beyond the BLS framework.

Third, our results are sensitive to prompt and parameter choices. We used a single, fixed prompt and default API settings to maintain comparability across occupations and models. Prior work shows that even small paraphrases can substantially shift representational outcomes [17]. The overall convergence of results across four models suggests these patterns likely reflect biases embedded in shared training data and processes, though we cannot rule out the possibility that they also stem from shared prompt sensitivities. Accordingly, our claims are bounded: we characterize cross-model patterns under one shared elicitation regime, not prompt-invariant behavior. To support cumulative work, we provide the exact prompt, field schema, and parsing code in the supplementary materials. Future audits should adopt multi-prompt designs (for example, 5–10 paraphrases spanning minor wording and discourse framing) and varied parameters (such as temperature or system instructions), then aggregate across prompts to estimate prompt-robust effects.

Fourth, we relied on a particular regression specification (centered-logit robust regression) to summarize systematic shift (α) and stereotype slope (β). This choice down-weights high-leverage points and improves interpretability, but alternative models—such as spline-based fits or employment-weighted regressions—might yield different magnitudes, especially at distributional extremes.

Finally, our scope was strictly limited to structured demographic attributes (gender and race/ethnicity). Other generated fields, such as names, motivations, biographies, and salaries, likely encode additional stereotypes and biases that warrant systematic analysis but are outside the scope of this study.

5.7 Future work

Promising directions include extending audits to intersectional, mixed race, and non-binary identities; examining representational bias in narrative attributes (e.g., motivations or biographies); and applying benchmarks from other national labor markets. In addition, adopting multi-prompt and multi-parameter strategies [17] would help assess

whether the patterns observed here are robust across elicitation conditions or primarily reflect shared prompt sensitivities across current model architectures.

6 Conclusion

Public concern over bias in generative AI has centered on how these systems reproduce and even intensify existing inequities. Our study contributes to this conversation by systematically analysing over 1.5 million generative AI occupational personas across 41 U.S. occupations. Our analysis of these occupational personas conclude general overrepresentation for women, Asian, and Hispanic workers, and underrepresentation for White and Black workers. Strong evidence suggests that generative AI models amplify demographic skews. Stereotypes are exaggerated for Hispanic, Asian, women, and White workers, while being reduced for Black workers through near erasure at typical occupational levels.

Tensions between demographic accuracy, social desirability, and perceived response quality make representational fairness a complex sociotechnical question. Continued work is needed in this space to further untangle the delicate interplay between these concepts to inform frameworks that hold generative systems accountable to widely shared values of equity and inclusion in the technologies that shape everyday life.

AI Usage Statement

In line with ACM's Policy on Authorship [5], we disclose that generative AI tools were used in this work. They were employed in two ways: (1) to generate the personas that were subsequently analyzed in our study (described in detail in the body of the paper), and (2) to assist with refining the presentation of results, including feedback on statistical analyses, language editing of author-written text, code editing of author-written scripts, and figure generation, where we used ChatGPT models (GPT-4 and GPT-5). All intellectual contributions, study design, analyses, interpretations, and claims in this paper are those of the human authors.

Acknowledgments

Ulrik Lyngs for feedback on figures. Sean Munson for feedback on idea. Matt Kay et al. for opensourcing their data, analyses, and code. Funding sources: Stitt Fellowship, SCU funding, etc.

References

[1] Rohaid Ali, Oliver Y Tang, Ian D Connolly, Hael A Abdulrazeq, Fatima N Mirza, Rachel K Lim, Benjamin R Johnston, Michael W Groff, Theresa Williamson, Konstantina Svokos, Tiffany J Libby, John H Shin, Ziya L Gokaslan, Curtis E Doberstein, James Zou, and Wael F Asaad. 2023. The face of a surgeon: An analysis of demographic representation in three leading artificial intelligence text-to-image generators. *bioRxiv* (May 2023), 2023.05.24.23290463. doi:10.1101/2023.05.24.23290463

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3290605.3300233

[3] Oghenemaro Anuyah, Ruyuan Wan, Cornelius Adejoro, Tom Yeh, Ronald Metoyer, and Karla Badillo-Urquiola. 2023. Cultural considerations in AI systems for the global south: A systematic review. In *Proceedings of the 4th African Human Computer Interaction Conference*. ACM, New York, NY, USA, 125–134. doi:10.1145/3628096.3629046

[4] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The silicon ceiling: Auditing GPT's race and gender biases in hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, New York, NY, USA, 1–18. doi:10.1145/3689904.3694699

[5] Association for Computing Machinery. 2023. ACM Policy on Authorship. https://www.acm.org/publications/policies/new-acm-policy-on-authorship. Approved by the ACM Publications Board on April 20, 2023. Accessed: 2025-09-11.

[6] Andrew C Baker, David F Larcker, Charles G McCLURE, Durgesh Saraph, and Edward M Watts. 2024. Diversity washing. *Journal of accounting research* 62, 5 (Dec. 2024), 1661–1709. doi:10.1111/1475-679x.12542

[7] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. 2021. Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA. doi:10.1145/3461702.3462610

[8] Eric P S Baumer. 2017. Toward human-centered algorithm design. *Big data & society* 4, 2 (Dec. 2017), 205395171771885. doi:10.1177/2053951717718854

[9] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 57, 1 (Jan. 1995), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

[10] Michael Bernstein, Angèle Christin, Jeffrey Hancock, Tatsunori Hashimoto, Chenyan Jia, Michelle Lam, Nicole Meister, Nathaniel Persily, Tiziano Piccardi, Martin Saveski, Jeanne Tsai, Johan Ugander, and Chunchen Xu. 2023. Embedding societal values into social media algorithms. *Journal of online trust & safety* 2, 1 (Sept. 2023). doi:10.54501/jots.v2i1.148

[11] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94, 4 (Sept. 2004), 991–1013. doi:10.1257/0002828042002561

[12] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 1493–1504. doi:10.1145/3593013.3594095

[13] Francine D Blau and Lawrence M Kahn. 2017. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature* 55, 3 (Sept. 2017), 789–865. doi:10.1257/jel.20160995

[14] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4356–4364. doi:10.5555/3157382.3157584

[15] Jane Castleman and Aleksandra Korolova. 2025. Aligned Generative Models Exhibit Adultification Bias. https://eclecticai.substack.com/p/aligned-generative-models-exhibit. Accessed: 2025-6-30.

[16] Christopher N Chapman and Russell P Milham. 2006. The personas' new clothes: Methodological and practical arguments against a popular method. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual Meeting* 50, 5 (Oct. 2006), 634–636. doi:10.1177/154193120605000503

[17] Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Rada Mihalcea, and Zhijing Jin. 2025. Causally testing gender bias in LLMs: A case study on occupational bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, Stroudsburg, PA, USA, 4984–5004. doi:10.18653/v1/2025.findings-naacl.281

[18] Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1504–1532. doi:10.18653/v1/2023.acl-long.84

[18] Sapna Cheryan, Sianna A Ziegler, Amanda K Montoya, and Lily Jiang. 2017. Why are some STEM fields more gender balanced than others? *Psychological bulletin* 143, 1 (Jan. 2017), 1–35. doi:10.1037/bul0000052

[19] Alan Cooper. 2004. *The inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity* (2 ed.). Sams Publishing, Indianapolis, IN. doi:10.5555/984201

[20] Nilanjana Dasgupta. 2011. Ingroup experts and peers as social vaccines who inoculate the self-concept: The stereotype inoculation model. *Psychological inquiry* 22, 4 (Oct. 2011), 231–246. doi:10.1080/1047840x.2011.607313

[21] Irune del Rio Gabiola. 2013. Globalizing the care chain: Representations of Latinas in maid in America. *Chasqui-revista De Literatura Latinoamericana* 42 (May 2013), 119–130. https://www.jstor.org/stable/43589516?casa_token=482AB__sJUQAAAAA:UqagNKlqQWme3zPexzUCCPkZSIjDZ9IwDYRT9ZwCRVe6CVtQ9nO_Ffpk7JaleY4NBZTvWby0v1xhmfFpX_CahRP4M_nYsY8D33DBuQGnKtDxkQaURNk

[22] Catherine D'Ignazio and Lauren F Klein. 2023. *Data feminism.* MIT Press, London, England. https://www.researchgate.net/profile/Emily-Yarrow-5/publication/363775476_BOOK_REVIEW_Data_Feminism_By_D'Ignazio_Catherine_and_Klein_Lauren_F_Cambridge_Massachusetts_and_London_England_The_MIT_Press_2020_ISBN_978-0-262-04400-4/links/63ea85744dcb750da757116a/BOOK-REVIEW-Data-Feminism-By-DIgnazio-Catherine-and-Klein-Lauren-F-Cambridge-Massachusetts-and-London-England-The-MIT-Press-2020-ISBN-978-0-262-04400-4.pdf

[23] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA. doi:10.1145/3301275.3302316

[24] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/2702123.2702556

[25] Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of AI-generated content: an examination of news produced by large language models. *Scientific reports* 14, 1 (March 2024), 5224. doi:10.1038/s41598-024-55686-2

[26] First Page Sage. 2025. *Top Generative AI Chatbots by Market Share – August 2025.* Technical Report. First Page Sage. https://firstpagesage.com/reports/top-generative-ai-chatbots/

[27] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM transactions on information systems* 14, 3 (July 1996), 330–347. doi:10.1145/230538.230561

[28] Future of Life Institute, Dylan Hadfield-Menell, Jessica Newman, Tegan Maharaj, Sneha Revanur, Stuart Russell, and David Krueger. 2025. *AI Safety Index: Summer 2025 Report.* Technical Report. Future of Life Institute. https://futureoflife.org/index

[29] Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How culture shapes what people want from AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15. doi:10.1145/3613904.3642660

[30] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Dec. 2021), 86–92. doi:10.1145/3458723

[31] R Stuart Geiger, Flynn O'Sullivan, Elsie Wang, and Jonathan Lo. 2025. Asking an AI for salary negotiation advice is a matter of concern: Controlled experimental perturbation of ChatGPT for protected and non-protected group discrimination on a contextual task with no clear ground truth answers. *PloS one* 20, 2 (Feb. 2025), e0318500. doi:10.1371/journal.pone.0318500

[32] Anna M Gorska and Dariusz Jemielniak. 2023. The invisible women: uncovering gender bias in AI-generated images of professionals. *Feminist media studies* 23, 8 (Nov. 2023), 4370–4375. doi:10.1080/14680777.2023.2263659

[33] Nico Grant. 2024. Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms. *The New York Times* (Feb. 2024). https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html

[34] J Brian Gray. 1994. Regression with graphics: A second course in applied statistics. *Technometrics: a journal of statistics for the physical, chemical, and engineering sciences* 36, 1 (Feb. 1994), 112–113. doi:10.1080/00401706.1994.10485408

[35] Douglas Guilbeault, Solène Delecourt, Tasker Hull, Bhargav Srinivasa Desikan, Mark Chu, and Ethan Nadler. 2024. Online images amplify gender bias. *Nature* 626, 8001 (Feb. 2024), 1049–1055. doi:10.1038/s41586-024-07068-x

[36] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3290605.3300830

[37] Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation gaps in machine learning practice. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA. doi:10.1145/3531146.3533233

[38] Hyejun Jeong, Shiqing Ma, and Amir Houmansadr. 2024. Bias similarity across Large Language Models. *arXiv [cs.LG]* (Oct. 2024). arXiv:2410.12010 [cs.LG] doi:10.48550/arXiv.2410.12010

[39] Soon-Gyo Jung, Joni Salminen, Kholoud Khalil Aldous, and Bernard J Jansen. 2025. PersonaCraft: Leveraging language models for data-driven persona development. *International journal of human-computer studies* 197, 103445 (March 2025), 103445. doi:10.1016/j.ijhcs.2025.103445

[40] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3313831.3376219

[41] M Kay, C Matuszek, and S A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd annual acm* (2015). doi:10.1145/2702123.2702520

[42] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D Procaccia. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on human-computer interaction* 3, CSCW (Nov. 2019), 1–35. doi:10.1145/3359283

[43] Sang Won Lee, Mary Morcos, Dong Won Lee, and Jason Young. 2024. Demographic representation of generative artificial intelligence images of physicians. *JAMA network open* 7, 8 (Aug. 2024), e2425993. doi:10.1001/jamanetworkopen.2024.25993

[44] Ang Li, Haozhe Chen, Hongseok Namkong, and Tianyi Peng. 2025. LLM Generated Persona is a Promise with a Catch. *arXiv [cs.CL]* (March 2025). arXiv:2503.16527 [cs.CL] doi:10.48550/arXiv.2503.16527

[45] Guoying Li. 2011. Robust Regression. In *Exploring Data Tables, Trends, and Shapes*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 281–343. doi:10.1002/9781118150702.ch8

[46] Penelope Lockwood. 2006. "someone like me can be successful": Do college students need same-gender role models? *Psychology of women quarterly* 30, 1 (March 2006), 36–46. doi:10.1111/j.1471-6402.2006.00260.x

[47] Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J Martin (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 48–55. doi:10.18653/v1/2021.nuse-1.5

[48] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large Language Models are Geographically Biased. *arXiv [cs.CL]* (Feb. 2024). arXiv:2402.02680 [cs.CL] doi:10.48550/arXiv.2402.02680

[49] Tara Matthews, Tejinder Judge, and Steve Whittaker. 2012. How do designers and user experience professionals actually perceive and use personas?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1219–1228. doi:10.1145/2207676.2208573

[50] Danaë Metaxa, Michelle A Gan, Su Goh, Jeff Hancock, and James A Landay. 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 (April 2021), 1–23. doi:10.1145/3449100

[51] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 220–229. doi:10.1145/3287560.3287596

[52] Lene Nielsen. 2004. *Engaging Personas and Narrative Scenarios*. Working Paper 6448. Copenhagen Business School, København, Denmark. https://research.cbs.dk/en/publications/engaging-personas-and-narrative-scenarios Final published version available via CBS Research Portal..

[53] Ihudiya Finda Ogbonnaya-Ogburu, Angela D R Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376392

[54] Maria Olsson and Sarah E Martiny. 2018. Does exposure to counterstereotypical role models influence girls' and women's gender stereotypes and career choices? A review of social psychological research. *Frontiers in psychology* 9 (Dec. 2018), 2264. doi:10.3389/fpsyg.2018.02264

[55] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA. doi:10.1145/3586183.3606763

[56] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3411764.3445315

[57] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA. doi:10.1145/3306618.3314244

[58] Joni Salminen, João M Santos, Soon-Gyo Jung, and Bernard J Jansen. 2024. Picturing the fictitious person: An exploratory study on the effect of images on user perceptions of AI-generated personas. *Computers in Human Behavior: Artificial Humans* 2, 1 (Jan. 2024), 100052. doi:10.1016/j.chbah.2024.100052

[59] Vanessa Sattele and Centro de Investigaciones de Diseño Industrial, UNAM. 2024. Generating user personas with AI: Reflecting on its implications for design. In *Proceedings of DRS*. Design Research Society. doi:10.21606/drs.2024.1024

[60] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on human-computer interaction* 3, CSCW (Nov. 2019), 1–33. doi:10.1145/3359246

[61] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on human-computer interaction* 4, CSCW1 (May 2020), 1–35. doi:10.1145/3392866

[62] Xiaoxiao Shang, Zhiyuan Peng, Qiming Yuan, Sabiq Khan, Lauren Xie, Yi Fang, and Subramaniam Vincent. 2022. DIANES: A DEI Audit Toolkit for News Sources. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3312–3317. doi:10.1145/3477495.3531660

[63] Nancy Signorielli and Aaron Bacue. 1999. Recognition and respect: A content analysis of prime-time television characters across three decades. *Sex roles* 40, 7-8 (April 1999), 527–544. doi:10.1023/a:1018883912900

[64] Vivek K Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 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 (Nov. 2020), 1281–1294. doi:10.1002/asi.24335

[65] Aleksandra Sorokovikova, Pavel Chizhov, Iuliia Eremenko, and Ivan P Yamshchikov. 2025. Surface fairness, deep bias: A comparative study of bias in language models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 206–227. doi:10.18653/v1/2025.gebnlp-1.20

[66] C Steele and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology* 69, 5 (Nov. 1995), 797–811. doi:10.1037/0022-3514.69.5.797

[67] Jane G Stout, Nilanjana Dasgupta, Matthew Hunsinger, and Melissa A McManus. 2011. STEMing the tide: using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). *Journal of personality and social psychology* 100, 2 (Feb. 2011), 255–270. doi:10.1037/a0021385

[68] U.S. Bureau of Labor Statistics. 2023. Current Population Survey: Annual Averages, 2023. https://www.bls.gov/cps/cps_aa2023.htm. Accessed: March 7, 2025.

[69] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao 'kenneth Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. *arXiv [cs.CL]* (Feb. 2023). arXiv:2302.02463 [cs.CL] doi:10.48550/arXiv.2302.02463

[70] James Vincent. 2025. We tried out DeepSeek. It works well — until we asked it about Tiananmen Square and Taiwan. https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan?utm_source=chatgpt.com. Accessed: 2025-9-10.

[71] Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of bias in Text-to-Image generation: Definition, evaluation, and mitigation. *arXiv [cs.CV]* (April 2024). arXiv:2404.01030 [cs.CV] doi:10.48550/arXiv.2404.01030

[72] Ruta Yemane and Mariña Fernández-Reino. 2021. Latinos in the United States and in Spain: the impact of ethnic group stereotypes on labour market outcomes. *Journal of ethnic and migration studies* 47, 6 (April 2021), 1240–1260. doi:10.1080/1369183x.2019.1622806

[73] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3290605.3300509

References

[1] Rohaid Ali, Oliver Y Tang, Ian D Connolly, Hael A Abdulrazeq, Fatima N Mirza, Rachel K Lim, Benjamin R Johnston, Michael W Groff, Theresa Williamson, Konstantina Svokos, Tiffany J Libby, John H Shin, Ziya L Gokaslan, Curtis E Doberstein, James Zou, and Wael F Asaad. 2023. The face of a surgeon: An analysis of demographic representation in three leading artificial intelligence text-to-image generators. *bioRxiv* (May 2023), 2023.05.24.23290463. doi:10.1101/2023.05.24.23290463

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3290605.3300233

[3] Oghenemaro Anuyah, Ruyuan Wan, Cornelius Adejoro, Tom Yeh, Ronald Metoyer, and Karla Badillo-Urquiola. 2023. Cultural considerations in AI systems for the global south: A systematic review. In *Proceedings of the 4th African Human Computer Interaction Conference*. ACM, New York, NY, USA, 125–134. doi:10.1145/3628096.3629046

[4] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The silicon ceiling: Auditing GPT's race and gender biases in hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, New York, NY, USA, 1–18. doi:10.1145/3689904.3694699

[5] Association for Computing Machinery. 2023. ACM Policy on Authorship. https://www.acm.org/publications/policies/new-acm-policy-on-authorship. Approved by the ACM Publications Board on April 20, 2023. Accessed: 2025-09-11.

[6] Andrew C Baker, David F Larcker, Charles G McCLURE, Durgesh Saraph, and Edward M Watts. 2024. Diversity washing. *Journal of accounting research* 62, 5 (Dec. 2024), 1661–1709. doi:10.1111/1475-679x.12542

[7] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. 2021. Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA. doi:10.1145/3461702.3462610

[8] Eric P S Baumer. 2017. Toward human-centered algorithm design. *Big data & society* 4, 2 (Dec. 2017), 205395171771885. doi:10.1177/2053951717718854

[9] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 57, 1 (Jan. 1995), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

[10] Michael Bernstein, Angèle Christin, Jeffrey Hancock, Tatsunori Hashimoto, Chenyan Jia, Michelle Lam, Nicole Meister, Nathaniel Persily, Tiziano Piccardi, Martin Saveski, Jeanne Tsai, Johan Ugander, and Chunchen Xu. 2023. Embedding societal values into social media algorithms. *Journal of online trust & safety* 2, 1 (Sept. 2023). doi:10.54501/jots.v2i1.148

[11] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94, 4 (Sept. 2004), 991–1013. doi:10.1257/0002828042002561

[12] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 1493–1504. doi:10.1145/3593013.3594095

[13] Francine D Blau and Lawrence M Kahn. 2017. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature* 55, 3 (Sept. 2017), 789–865. doi:10.1257/jel.20160995

[14] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 4356–4364. doi:10.5555/3157382.3157584

[15] Jane Castleman and Aleksandra Korolova. 2025. Aligned Generative Models Exhibit Adultification Bias. https://eclecticai.substack.com/p/aligned-generative-models-exhibit. Accessed: 2025-6-30.

[16] Christopher N Chapman and Russell P Milham. 2006. The personas' new clothes: Methodological and practical arguments against a popular method. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual Meeting* 50, 5 (Oct. 2006), 634–636. doi:10.1177/154193120605000503

[17] Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Rada Mihalcea, and Zhijing Jin. 2025. Causally testing gender bias in LLMs: A case study on occupational bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, Stroudsburg, PA, USA, 4984–5004. doi:10.18653/v1/2025.findings-naacl.281

[18] Sapna Cheryan, Sianna A Ziegler, Amanda K Montoya, and Lily Jiang. 2017. Why are some STEM fields more gender balanced than others? *Psychological bulletin* 143, 1 (Jan. 2017), 1–35. doi:10.1037/bul0000052

[19] Alan Cooper. 2004. *The inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity* (2 ed.). Sams Publishing, Indianapolis, IN. doi:10.5555/984201

[20] Nilanjana Dasgupta. 2011. Ingroup experts and peers as social vaccines who inoculate the self-concept: The stereotype inoculation model. *Psychological inquiry* 22, 4 (Oct. 2011), 231–246. doi:10.1080/1047840x.2011.607313

[21] Irune del Rio Gabiola. 2013. Globalizing the care chain: Representations of Latinas in maid in America. *Chasqui-revista De Literatura Latinoamericana* 42 (May 2013), 119–130. https://www.jstor.org/stable/43589516?casa_token=482AB__sJUQAAAAA:UqagNKlqQWme3zPexzUCCPkZSIjDZ9IwDYRT9ZwCRVe6CVtQ9nO_Ffpk7JaleY4NBZTvWby0v1xhmfFpX_CahRP4M_nYsY8D33DBuQGnKtDxkQaURNk

[22] Catherine D'Ignazio and Lauren F Klein. 2023. *Data feminism.* MIT Press, London, England. https://www.researchgate.net/profile/Emily-Yarrow-5/publication/363775476_BOOK_REVIEW_Data_Feminism_By_D'Ignazio_Catherine_and_Klein_Lauren_F_Cambridge_Massachusetts_and_London_England_The_MIT_Press_2020_ISBN_978-0-262-04400-4/links/63ea85744dcb750da757116a/BOOK-REVIEW-Data-Feminism-By-DIgnazio-Catherine-and-Klein-Lauren-F-Cambridge-Massachusetts-and-London-England-The-MIT-Press-2020-ISBN-978-0-262-04400-4.pdf

[23] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA. doi:10.1145/3301275.3302316

[24] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/2702123.2702556

[25] Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of AI-generated content: an examination of news produced by large language models. *Scientific reports* 14, 1 (March 2024), 5224. doi:10.1038/s41598-024-55686-2

[26] First Page Sage. 2025. *Top Generative AI Chatbots by Market Share – August 2025.* Technical Report. First Page Sage. https://firstpagesage.com/reports/top-generative-ai-chatbots/

[27] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM transactions on information systems* 14, 3 (July 1996), 330–347. doi:10.1145/230538.230561

[28] Future of Life Institute, Dylan Hadfield-Menell, Jessica Newman, Tegan Maharaj, Sneha Revanur, Stuart Russell, and David Krueger. 2025. *AI Safety Index: Summer 2025 Report.* Technical Report. Future of Life Institute. https://futureoflife.org/index

[29] Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How culture shapes what people want from AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15. doi:10.1145/3613904.3642660

[30] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Dec. 2021), 86–92. doi:10.1145/3458723

[31] R Stuart Geiger, Flynn O'Sullivan, Elsie Wang, and Jonathan Lo. 2025. Asking an AI for salary negotiation advice is a matter of concern: Controlled experimental perturbation of ChatGPT for protected and non-protected group discrimination on a contextual task with no clear ground truth answers. *PloS one* 20, 2 (Feb. 2025), e0318500. doi:10.1371/journal.pone.0318500

[32] Anna M Gorska and Dariusz Jemielniak. 2023. The invisible women: uncovering gender bias in AI-generated images of professionals. *Feminist media studies* 23, 8 (Nov. 2023), 4370–4375. doi:10.1080/14680777.2023.2263659

[33] Nico Grant. 2024. Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms. *The New York Times* (Feb. 2024). https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html

[34] J Brian Gray. 1994. Regression with graphics: A second course in applied statistics. *Technometrics: a journal of statistics for the physical, chemical, and engineering sciences* 36, 1 (Feb. 1994), 112–113. doi:10.1080/00401706.1994.10485408

[35] Douglas Guilbeault, Solène Delecourt, Tasker Hull, Bhargav Srinivasa Desikan, Mark Chu, and Ethan Nadler. 2024. Online images amplify gender bias. *Nature* 626, 8001 (Feb. 2024), 1049–1055. doi:10.1038/s41586-024-07068-x

[36] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3290605.3300830

[37] Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation gaps in machine learning practice. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA. doi:10.1145/3531146.3533233

[38] Hyejun Jeong, Shiqing Ma, and Amir Houmansadr. 2024. Bias similarity across Large Language Models. *arXiv [cs.LG]* (Oct. 2024). arXiv:2410.12010 [cs.LG] doi:10.48550/arXiv.2410.12010

[39] Soon-Gyo Jung, Joni Salminen, Kholoud Khalil Aldous, and Bernard J Jansen. 2025. PersonaCraft: Leveraging language models for data-driven persona development. *International journal of human-computer studies* 197, 103445 (March 2025), 103445. doi:10.1016/j.ijhcs.2025.103445

[40] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3313831.3376219

[41] M Kay, C Matuszek, and S A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd annual acm* (2015). doi:10.1145/2702123.2702520

[42] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D Procaccia. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on human-computer interaction* 3, CSCW (Nov. 2019), 1–35. doi:10.1145/3359283

[43] Sang Won Lee, Mary Morcos, Dong Won Lee, and Jason Young. 2024. Demographic representation of generative artificial intelligence images of physicians. *JAMA network open* 7, 8 (Aug. 2024), e2425993. doi:10.1001/jamanetworkopen.2024.25993

[44] Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. LLM Generated Persona is a Promise with a Catch. *arXiv [cs.CL]* (March 2025). arXiv:2503.16527 [cs.CL] doi:10.48550/arXiv.2503.16527

[45] Guoying Li. 2011. Robust Regression. In *Exploring Data Tables, Trends, and Shapes*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 281–343. doi:10.1002/9781118150702.ch8

[46] Penelope Lockwood. 2006. "someone like me can be successful": Do college students need same-gender role models? *Psychology of women quarterly* 30, 1 (March 2006), 36–46. doi:10.1111/j.1471-6402.2006.00260.x

[47] Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J Martin (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 48–55. doi:10.18653/v1/2021.nuse-1.5

[48] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large Language Models are Geographically Biased. *arXiv [cs.CL]* (Feb. 2024). arXiv:2402.02680 [cs.CL] doi:10.48550/arXiv.2402.02680

[49] Tara Matthews, Tejinder Judge, and Steve Whittaker. 2012. How do designers and user experience professionals actually perceive and use personas?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1219–1228. doi:10.1145/2207676.2208573

[50] Danaë Metaxa, Michelle A Gan, Su Goh, Jeff Hancock, and James A Landay. 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 (April 2021), 1–23. doi:10.1145/3449100

[51] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 220–229. doi:10.1145/3287560.3287596

[52] Lene Nielsen. 2004. *Engaging Personas and Narrative Scenarios*. Working Paper 6448. Copenhagen Business School, København, Denmark. https://research.cbs.dk/en/publications/engaging-personas-and-narrative-scenarios Final published version available via CBS Research Portal..

[53] Ihudiya Finda Ogbonnaya-Ogburu, Angela D R Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376392

[54] Maria Olsson and Sarah E Martiny. 2018. Does exposure to counterstereotypical role models influence girls' and women's gender stereotypes and career choices? A review of social psychological research. *Frontiers in psychology* 9 (Dec. 2018), 2264. doi:10.3389/fpsyg.2018.02264

[55] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA. doi:10.1145/3586183.3606763

[56] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3411764.3445315

[57] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA. doi:10.1145/3306618.3314244

[58] Joni Salminen, João M Santos, Soon-Gyo Jung, and Bernard J Jansen. 2024. Picturing the fictitious person: An exploratory study on the effect of images on user perceptions of AI-generated personas. *Computers in Human Behavior: Artificial Humans* 2, 1 (Jan. 2024), 100052. doi:10.1016/j.chbah.2024.100052

[59] Vanessa Sattele and Centro de Investigaciones de Diseño Industrial, UNAM. 2024. Generating user personas with AI: Reflecting on its implications for design. In *Proceedings of DRS*. Design Research Society. doi:10.21606/drs.2024.1024

[60] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on human-computer interaction* 3, CSCW (Nov. 2019), 1–33. doi:10.1145/3359246

[61] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on human-computer interaction* 4, CSCW1 (May 2020), 1–35. doi:10.1145/3392866

[62] Xiaoxiao Shang, Zhiyuan Peng, Qiming Yuan, Sabiq Khan, Lauren Xie, Yi Fang, and Subramaniam Vincent. 2022. DIANES: A DEI Audit Toolkit for News Sources. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3312–3317. doi:10.1145/3477495.3531660

[63] Nancy Signorielli and Aaron Bacue. 1999. Recognition and respect: A content analysis of prime-time television characters across three decades. *Sex roles* 40, 7-8 (April 1999), 527–544. doi:10.1023/a:1018883912900

[64] Vivek K Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. 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 (Nov. 2020), 1281–1294. doi:10.1002/asi.24335

[65] Aleksandra Sorokovikova, Pavel Chizhov, Iuliia Eremenko, and Ivan P Yamshchikov. 2025. Surface fairness, deep bias: A comparative study of bias in language models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 206–227. doi:10.18653/v1/2025.gebnlp-1.20

[66] C Steele and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology* 69, 5 (Nov. 1995), 797–811. doi:10.1037/0022-3514.69.5.797

[67] Jane G Stout, Nilanjana Dasgupta, Matthew Hunsinger, and Melissa A McManus. 2011. STEMing the tide: using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). *Journal of personality and social psychology* 100, 2 (Feb. 2011), 255–270. doi:10.1037/a0021385

This work is currently under peer review for publication. This version is a preprint and has not undergone formal peer review.

[68] U.S. Bureau of Labor Statistics. 2023. Current Population Survey: Annual Averages, 2023. https://www.bls.gov/cps/cps_aa2023.htm. Accessed: March 7, 2025.

[69] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao 'kenneth Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. *arXiv [cs.CL]* (Feb. 2023). arXiv:2302.02463 [cs.CL] doi:10.48550/arXiv.2302.02463

[70] James Vincent. 2025. We tried out DeepSeek. It works well — until we asked it about Tiananmen Square and Taiwan. https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan?utm_source=chatgpt.com. Accessed: 2025-9-10.

[71] Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of bias in Text-to-Image generation: Definition, evaluation, and mitigation. *arXiv [cs.CV]* (April 2024). arXiv:2404.01030 [cs.CV] doi:10.48550/arXiv.2404.01030

[72] Ruta Yemane and Mariña Fernández-Reino. 2021. Latinos in the United States and in Spain: the impact of ethnic group stereotypes on labour market outcomes. *Journal of ethnic and migration studies* 47, 6 (April 2021), 1240–1260. doi:10.1080/1369183x.2019.1622806

[73] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3290605.3300509

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009