How nouns and adjectives contribute to perception of gender stereotypes: Comparing humans, LLMs and AI image generators

Elsi Kaiser (emkaiser@usc.edu) & Ashley Adji, University of Southern California

Both humans and large language models (LLMs) exhibit gender biases [3,5]. But it is unclear to what extent humans' and LLMs' biases align, how different expressions (e.g. nouns, adjectives) contribute, and to what extent biases extend to AI image-generators (e.g. DALL-E [1,2]). We use pronouns to test how adjectives and nouns guide humans' activation of gender stereotypes, to test *human reference resolution and stereotypes*. We compare humans, LLMs and image generators to *understand AI biases*, as a step toward mitigating their consequences.

Exp.1-2 test how gender-biased role nouns (e.g. *plumber*, *nurse*) and gender-biased adjectives (e.g. *powerful*, *kind*) influence humans' and GPT's assumptions about gender in a fill-in-the-blank task in English. Exp.3 tests how these words influence images DALL-E creates.

Exp.1-2: Text-based task. In **Exp.1,** 50 native U.S.-English speakers filled in blanks in sentences whose subjects were male-biased, female-biased or neutral role nouns. On targets, the blanks were designed to elicit pronouns (*her, his, their,* see ex.1-2) to reveal participants' inferences / assumptions about the gender of the subject noun. The study included 27 targets and 33 fillers. In **Exp.2**, GPT-4o completed the same task (version 4o, with 'memory' set to off and 'temporary chat' set to on, to help avoid learning effects and priming), 90 times.

Design (3x3 Latin Square): We manipulated whether the description of the critical referent includes (i) a male-biased, female-biased adjective, or no adjective, and whether the role noun was (ii) male-biased, female-biased, or neutral (ex.1-2, Table 1, roles and adjectives were selected based on norms from [4,6]). We test three hypotheses:

Hypothesis 1 Grammar-based asymmetry: Grammatical category (N vs. Adj) determines which information source dominates. noun information is privileged over adjectival information, *or* vice versa: Either nouns or adjectives trigger inferences about referent gender.

Hyp 2 Gender-based asymmetry: One gender is privileged (e.g. 'assume female if any cue suggests female'), regardless of grammatical encoding (i.e., whether info is on N or Adj).

Hyp 3 Symmetrical combination: Nominal and adjectival information, and male and female biases, have equal effects: When cues conflict (MaleAdj+FemaleN; FemaleAdj+MaleN), 'his' and 'her' should occur equally; if both point to one gender, pronouns for that gender occur more.

Human results (Fig.1) show a grammar-based asymmetry (Hyp 1). With *both* female- and male-biased nouns, regardless of adjective, people tend to produce *noun-gender-matching pronouns* (glmer, above chance, p's<.02). Adjective effects only emerge with neutral nouns (and only male-biased adjectives, p<.018). Overall, humans prioritize gender cues carried by *nouns*.

GPT-4o results (Fig.2) exhibit a gender-based asymmetry, in line with Hyp 2. With *female*-biased nouns, regardless of adjective, 'her' is preferred (above chance, p's<.03). With *female*-biased adjectives, neutral nouns prefer 'her' (above chance, p<.02) and even male nouns no longer prefer 'his' ('his/her' at chance, p's>0.3). **As a whole**: both humans and GPT-4o are susceptible to gender bias, but humans prioritize gender cues on nouns over adjectives, while GPT-4o seems sensitive to any hint of female bias, independent of its grammatical status.

Exp.3: What about **image generation**? We used DALL-E 3 (May/June 2024 paid version) to generate >300 images using prompts containing (i) a male-biased or female-biased adjective or no adjective, and (ii) male-biased or female-biased role nouns or the noun 'person' (ex.3). We used simplistic prompts (e.g. "two images: the tough gymnast" to avoid any syntactic effects).

DALL-E results (Fig.3) hint at a symmetrical combination (Hyp 3). For female- and male-biased role nouns, DALL-E mostly generated people matching nouns' gender bias; surprisingly, it treats 'person' as male. Crucially, conflicting adjectives weaken or eliminate this effect.

Disentangling effects of gender-stereotypical nouns vs. adjectives reveals (i) humans exhibit grammar-based asymmetries in gender-bias effects, and (ii) humans and LLMs are both biased but diverge in unexpected ways, highlighting the value of systematically testing sources of bias.

(1) ***Sample item Exp.1-2*** *(27 male- & female-biased roles and 27 male- & female-biased adj)*

Many people were trying to talk at once…	But the { **ø / nice$_{FEM}$ / greedy$_{MALE}$**} **sales assistant$_{FEM}$** kept ___ mouth shut.
	But the { **ø / nice$_{FEM}$ / greedy$_{MALE}$** } **bus driver$_{MALE}$** kept ___ mouth shut.
	But the { **ø / nice$_{FEM}$ / greedy$_{MALE}$** } **musician$_{NEUT}$** kept ___ mouth shut.

(2) ***Samples of other frames*** *(27 frames used; participants typed in a word for the ___ blank)*
(a) All of a sudden, the [critical referent] heard a noise that attracted ___ attention.
(b) Sometimes it is best to not get involved, so the [crit ref] decided to mind ___ own business.

Male-biased role nouns	*plumber, boxer, butcher, mechanic, farmer, etc*
Female-biased role nouns	*florist, nanny, wedding planner, secretary, teacher, etc*
Neutral role nouns	*editor, photographer, writer, tour guide, proofreader, etc*
Male-biased adjectives	*powerful, dangerous, handsome, tough, wealthy, etc*
Female-biased adjectives	*kind, sweet, gorgeous, sentimental, graceful, etc*

Table 1. Examples of female- and male-biased role nouns and adjectives (selected to be as male- and female-biased as possible from norms by Misersky et al. 2014 and Scott et al. 2019)

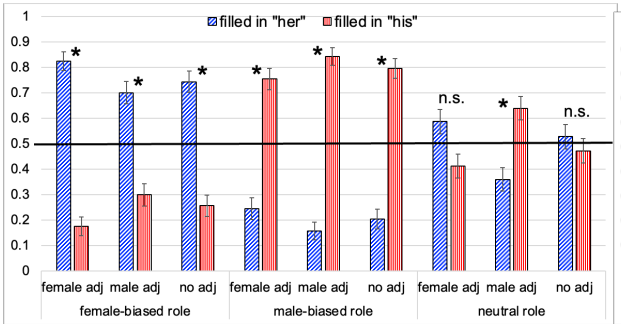


Fig.1. Exp.1: What pronoun did **people** type in the blank? * = differs from chance. ('their' was used equally in all conditions, not shown.)

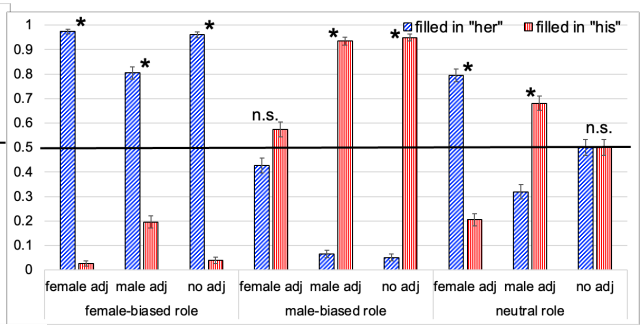


Fig.2. Exp.2: What pronoun did **GPT-4o** fill in the blank? * = differs from chance. ('their' use not shown, only occurred 1.4% of the time.)

(3) **Some examples** showing wording entered into DALL-E (2 images/prompt, system's default)
(a) two images: the tough gymnast [male-biased adjective + female-biased role noun]
(b) two images: the tough carpenter [male-biased adjective + male-biased role noun]
(c) two images: the tough person [male-biased adjective + 'person']

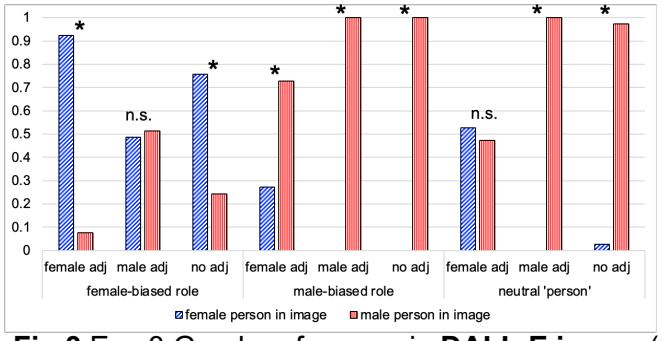
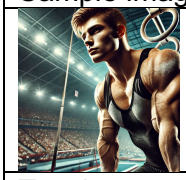


Fig.3 Exp.3 Gender of person in **DALL-E image** (fig is proportions; statistics done on counts with one-way chi-sq. * = # of male vs. female images differs)

Sample images from DALL-E


Tough gymnast


Lovely gymnast


Emotional nanny


Powerful nanny

[1]Ananya'24 *AI image generators often give racist and sexist results* [2]Bianchi etal'23 *Easily accessible text-to-image generation* [3]Borji'23 *Categorical archive of chatgpt failures* [4] Misersky etal'14 *Norms on the gender perception of role nouns* [5]Doshi et al'23 *ChatGPT: temptations of progress* [6]Scott etal'19 *Glasgow Norms* [7]Walther etal'24 *Gendered nature of AI*