2025. Proc Ling Soc Amer 10(1). 5954. https://doi.org/10.3765/plsa.v10i1.5954.

On gender stereotypicality in nouns and adjectives: Comparing humans, large language models and text-to-image generators

Elsi Kaiser & Ashley Adji*

Abstract. Both humans and large language models (LLMs) are known to exhibit effects of gender stereotypicality. We conducted a series of studies to systematically assess to what extent humans' and LLMs' interpretational patterns align, how different kinds of linguistic expressions (role nouns vs. adjectives) contribute, and to what extent these patterns extend to text-to-image models. Experiments 1 and 2 test how gender-biased role nouns (e.g. *plumber*, *nurse*) and adjectives (e.g. *powerful*, *kind*) influence humans' and GPT-4o's assumptions about gender in a fill-in-the-blank task. Experiment 3 tests how role nouns and adjectives influence images created by the image generator DALL-E 3 (a text-to-image model). Our results show that humans, LLMs and text-to-image models' outputs are all influenced by gender stereotypes but diverge in unexpected ways.

Keywords. gender stereotypes; gender bias; pronoun production; experimental linguistics; sociolinguistics; large language models; artificial intelligence; text-to-image generators; role nouns; adjectives

1. Introduction. Both humans and large language models (LLMs, e.g. GPT, Claude, Llama) are known to exhibit sensitivity to gender stereotypicality (e.g. Vig et al. 2020, Kotek et al. 2023, Doshi et al. 2023, Wan et al. 2023, Dong et al. 2024, Zhao et al. 2024, and many others). For example, in a recent paper, Kotek et al. (2023) tested a variety of LLMs and found that LLMs make systematic assumptions about the genders associated with different professions. They used prompts such as (1) and manipulated whether the pronoun was *he* or *she* (see also Marin & Eger 2024). Crucially, this pronoun is in principle ambiguous and could refer to the doctor or to the nurse. Thus, the key question is whether the LLMs' interpretation of who the pronoun picks out is influenced by the gender stereotypicality of the two professions mentioned in the sentence.

(1) In the sentence: "The doctor phoned the nurse because {she / he was} late for the morning shift", who was late for the morning shift?

Kotek et al. found that when the pronoun was *he*, LLMs chose the doctor as the antecedent 63-66% of the time (depending on LLM model). But when the pronoun was *she*, LLMs chose the nurse as the antecedent 74-78% of the time (depending on LLM model). Thus, LLMs exhibit a clear bias to interpret pronouns in line with the gender stereotypes associated with different professions. (We follow Misersky et al. (2014) and use the term 'gender stereotypicality' to refer to "generalized beliefs or expectations about whether a specific (social or occupational) role is more likely to be held by one gender or the other" (Misersky et al. 2014: 842).)

Text-to-image (T2I) models, e.g. OpenAI's DALL-E 3 and Google's Gemini, are also susceptible to gender stereotypes (e.g. Bianchi et al. 2023, Naik & Nushi 2023, Wan et al. 2024, Girrbach et al. 2025; Friedrich et al. 2024 for a multilingual perspective). For example, Girrbach et. al. (2025) conducted a large-scale analysis of >2,000,000 images generated by T2I models,

* We gratefully acknowledge funding from the USC Summer Undergraduate Research Fund. Many thanks to the audience at the 2025 LSA Annual Meeting for helpful comments and feedback. Authors: Elsi Kaiser, University of Southern California (emkaiser@usc.edu) & Ashley Adji, University of Southern California (adji@usc.edu).

and found effects of gender-stereotypical information with T2I models' depictions of people engaged in a range of activities (e.g. *gaming* vs. *shopping*) and people with various occupations (e.g *carpenter* vs. *dental hygienist*). Fraser et al. (2023) found that adjectives selected from work on social cognition – e.g. looking at notions such as agency, with adjectives like *competitive* vs. *passive* – have a systematic effect on the kinds of people generated by the Midjourney model (though DALL-E 2 showed no effects for agency-related adjectives). For recent benchmarking and debiasing work on gender bias, see e.g. Luo et al. (2024) and Li et al. (2025b).

When assessing T2I models, it is important to keep in mind that, as noted by Wan et al. (2024), "since only gender presentation and roles may be perceived from model-synthesized images, the concept of "gender" in these studies refers to perceived gender presentation and roles, not gender or sexual identity" (Wan et al. 2024. P.3). See also Fraser et al. (2023) for related discussion on the ethical considerations involved in labeling people based on gender.

In the present work, we seek to further understanding of how humans' assumptions about gender stereotypes – stemming not only from role nouns but also adjectives – compare to those of LLMs and text-to-image (T2I) generators. To obtain a comparable data set, we conducted a study with (i) humans and (ii) GPT-4o, using the same task and the same items. We used an 'implicit' method that did not explicitly mention 'gender' or 'stereotype.' We also tested the image generator DALL-E 3, using a modified version of the task.

Crucially, in addition to probing how professional role nouns elicit assumptions about the referent's gender, we concurrently test for potential effects of gender-biased adjectives that describe personality traits and emotional states (e.g. *delightful, angry, tough, kind*). To the best of our knowledge, although there is a lot of research on role nouns/professions, prior work comparing humans and language models has not systematically explored the consequences of potentially conflicting cues conveyed by role nouns vs. adjectives (e.g. *a kind doctor* vs. *a tough doctor*). Given the importance of adjectives in letters of recommendation, performance evaluations and other contexts, a better understanding of their usage patterns is valuable. Furthermore, by testing adjectives and role nouns, we can gain insights into how humans and generative AI handle 'cue conflict' situations, i.e., situations where two cues diverge from each other.

In what follows, we report three experiments. Experiment 1 is a fill-in-the-blank study with human participants. Experiment 2 uses the same items and the same language-based task, with GPT-4o. Experiment 3 probes the DALL-E 3 image generator, using a slightly modified version of the task (adjusted to the realm of images). In all three experiments, we test gender-biased role nouns (e.g. *nurse, mechanic*) and gender-biased adjectives (e.g. *bubbly, tough*) – identified using pre-existing norms – in configurations where the biases align/point in the same direction (e.g. (2a,2c)) as well as situations where they diverge/conflict (e.g. (2b,d)), to assess how they contribute to assumptions about the referent's gender.

(2) a. the angry golfer *[male-biased adjective + male-biased role noun]*
b. the delightful golfer *[female-biased adjective + male-biased role noun]*
c. the angry make-up artist *[male-biased adjective + female-biased role noun]*
d. the delightful make-up artist *[female-biased adjective + female-biased role noun]*

1.2 GENDER BIAS FROM ROLE NOUNS AND ADJECTIVES. We consider three ways in which the information from role nouns and adjectives could shape the inferences that people (and AI models) make about the gender of the referent: (i) the Information-type Asymmetry hypothesis, (ii) the Information-type Symmetry hypothesis and (iii) the Gender Asymmetry hypothesis.

It is important to note that in the current work, the distinction between (a) nouns and (b) ad-

jectives is correlated with the distinction between (a) a profession/job and (b) a characteristic / property / emotional state of a person, because the former is consistently expressed by a noun and the latter by an adjective. Thus, our main aim is not to make strong claims about whether it's specific parts of speech (nouns/adjectives) *or* specific kinds of information (professions vs. traits / characteristics of humans) that have a bigger influence on people's assumptions about referent gender. This is because the noun-vs-adjective distinction is (necessarily) 'confounded' with the profession-vs-characteristic distinction. Nevertheless, given that prior work – albeit on topics other than professions and gender – suggests that nouns trigger stereotypical inferences more than adjectives (e.g. Carnaghi et al. 2008), we want to test if the gender associations of role nouns have a stronger effect than those of adjectives.

Crucially, we want to test what happens when different sources of stereotypical information that can trigger inferences about gender *converge* (when the role noun and the adjective point in the same direction) vs. *divergence* (when the role noun and the adjective differ in stereotypical gender association). In particular, what happens in cases of divergence – which can be characterized as 'cue conflict' situations: which cue is more influential? To test this, we needed a configuration where there are two linguistic elements that can be independently varied to carry information about gender stereotypes, and so we chose to test nouns and adjectives.

How does information about stereotypical gender carried by nouns and adjectives interact? In the present paper we focus on three possibilities:

First, according to the **Information-type Asymmetry Hypothesis**, the type of information determines how influential it is in guiding inferences about gender. By 'type', we mean information carried by role nouns (i.e., the person's profession) vs. information carried by adjectives (i.e., information about the person's personality traits or emotional state). For ease of exposition, we refer to these as cues carried by/encoded by/stemming from *role nouns* vs. *adjectives*, but remind the reader that part of speech is correlated with information type as discussed above.

According to this hypothesis, gender cues stemming from role nouns vs. and those stemming from adjectives differ in how strongly they guides comprehenders' inferences about referent gender. This hypothesis predicts that when the two kinds of information conflict (e.g. the role noun points to male and the adjective points to female or vice versa), participants will give more weight to one information source than the other, rather than treating the referent as equally likely to be female or male. In other words, under this view, cue conflict does not result a 50-50 outcome. Rather, situations where the cues conflict will reveal which source of gender information is more powerful: role nouns or adjectives.

Alternatively, according to the **Information-type Symmetry Hypothesis**, information of both types – from role nouns and adjectives – is equally weighted. Thus, under this view, when the cues diverge, e.g. if a male-biased adjective is combined with a female-biased role noun or vice versa, comprehenders should be equally likely to assume the individual is male or female.

Finally, under the **Gender Asymmetry Hypothesis**, one gender is privileged such that any cues towards this gender, whether coming from the noun or the adjective, will be weighted more heavily than cues in favor of the other gender. Thus, in an extreme form, this hypothesis could be paraphrased as 'assume female if any cue suggests female' or as 'assume male if any cue suggests male,' regardless of whether the information is on the noun or adjective. Situations where the gender cues diverge can reveal whether there exists a gender that is privileged in terms of how much it influences assumptions about gender.

2. Experiment 1: Language task by humans. Experiment 1 used a fill-in-the-blank task to test how descriptions of people with matching vs. conflicting gender cues coming from role nouns

and adjectives influences assumptions about the gender of the person being described.

2.1 METHOD

2.1.1. PARTICIPANTS. 55 native U.S. English speakers, recruited via Prolific, participated remotely over the internet. Five participants were excluded for not being native English speakers born in the U.S., which left 50 participants for the final analysis.

2.1.2 MATERIALS AND DESIGN. Target sentences were of the form shown in (3). In a 3x3 design, we manipulated the gender bias of the role noun and the adjective. More specifically, 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. (We did not include a condition with neutral adjectives due to difficulty identifying a sufficient number of neutral adjectives.) The 27 targets were constructed using 27 different male- and female-biased role nouns, 27 gender-balanced (neutral) role nouns, as well as 27 different male- and female-biased adjectives. The adjectives and role nouns were selected using existing large-scale norms from Misersky et al. (2014) and Scott et al. (2019), to ensure that they elicit the intended gender associations. We discuss the selection criteria in more detail below. In (3), for ease of exposition, subscripts (M for male-biased, F for female-biased and N for neutral) indicate the bias of each role noun and adjective.

Each experimental item contains a blank that participants fill in with one word. The blanks on target trials were designed so that they could be felicitously filled in with a possessive pronoun (*his, her, their*, see (3)). Additional examples of targets are in (4). The experiment included 27 targets, with 27 different nouns and adjectives. Targets were presented to participants using a Latin-Square design, so that a particular participant only saw each target once.

(3) a. *Example: Female-biased profession*
Many people were trying to talk at once. But the {ø / nice_F / greedy_M } sales assistant_F kept ___ mouth shut.
b. *Example: Male-biased profession*
Many people were trying to talk at once. But the {ø / nice_F / greedy_M } bus driver_M kept ___ mouth shut.
c. *Example: Neutral profession*
Many people were trying to talk at once. But the {ø / nice_F / greedy_M } musician_N kept ___ mouth shut.

(4) 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 [critical referent] decided to mind ___ own business.

Participants were instructed to fill in each blank with one word. In addition to 27 targets, the study included 33 fillers. Fillers had blanks that could be filled in with a variety of other kinds of words, including prepositions, adjectives, and nouns. Examples of fillers are in (5).

(5) a. The sun was shining. The playful toddler saw a bird ___ the big window.
b. The zombie was disguised to seem human; a ___ in sheep's clothing, as they say.
c. During spring break, the college student watched a new episode ___ Netflix every night.

This way, we could test people's assumptions about the gender of the critical referents on target trials *without* having to use questions that directly reveal that we interested in assumptions about gender. On targets, the pronouns participants type in reveal what inferences / assumptions

they were making about the referent's gender. After the main experiment, participants also filled in a brief questionnaire about their own gender attitudes; those results are not discussed here.

Let us now consider how adjectives and role nouns were selected. Using existing norms, we chose the adjectives and role nouns to be as clearly associated with stereotypically male and female referents as possible (with neutral role nouns balanced between male and female).

Selection of role nouns. The role nouns were selected using norms collected by Misersky et al. (2014). Misersky et al. asked participants to estimate "the extent to which the presented social and occupational groups actually consisted of women and men" on a scale of "100 men, 0 women" to "100 women, 0 men" – in other words, to estimate the ratio of women vs. men in a given group (e.g. bankers, cyclists etc). The norms report this score as a proportion from 0 (all men) to 1 (all women) – i.e., the bigger the number, the higher the proportion of women. (We intentionally use norms based on humans' impressions, not labor statistics, as these do not always align.)

The 27 neutral role nouns we selected have a mean rating of 0.493 (SD 0.03) and range from 0.45 ('sculptors') to 0.55 ('interpreters', 'psychiatrists'). The 27 male-biased nouns have a mean of 0.236 (SD 0.04) and range from 0.16 ('miners') to 0.35 ('technicians'). The 27 female-biased nouns have a mean of 0.718 (SD 0.06) and range from 0.62 ('violinists') to 0.84 ('beauticians').

Unpaired t-tests confirm that, as intended, the three sets of nouns that we selected differ significantly in their gender estimates in the intended way. The female-biased role nouns differ from both the neutral and male-biased role nouns: The female-biased role that we selected are rated as having a significantly higher proportion of women (and lower proportion of men) than the neutral roles nouns ($t(52)=17.975$, $p<.0001$) and the male-biased role nouns ($t(52)=35.802$, $p<.0001$). The neutral role nouns also differ significantly from the male-biased role nouns in the intended direction (rated as having a lower proportion of men; $t(52)=26.867$, $p<.0001$). Examples of some of the selected role nouns are given in Table 1.

Selection of adjectives. The adjectives were selected using the norms from Scott et al. (2019), with the aim of identifying maximally female-biased/female-associated adjectives and maximally male-biased/male-associated adjectives. Scott et al. asked participants to rate words on a seven-point scale from 'very feminine' (1) to 'very masculine' (7).

The 27 female-biased adjectives that we selected have an average rating of 2.24 (SD 0.42), with a range from 1.441 ('beautiful') to 2.8 ('delightful'). The 27 male-biased adjectives have an average rating of 5.28 (SD 0.41), with a range from 4.8 ('wealthy') to 6.171 ('handsome'). Unpaired t-tests confirm that these two sets of adjectives differ significantly in their ratings in the intended way ($t(52)=27.037$, $p<.0001$). Examples of some of the adjectives are in Table 1. Many of the adjectives that have stereotypical gender associations are subjective and can be classified as predicates of personal taste or multidimensional adjectives (see e.g. Lasersohn 2005, Sassoon 2013, Kennedy 2013, Stojanovic & McNally 2017, Kaiser & Herron Lee 2018).

If we convert the adjectives' 7-point scale ratings from Scott et al. to a proportion from 0 (all men) to 1 (all women), to render them comparable to the Misersky et al. role noun norms, the numbers for nouns and adjectives very similar – which is what we intended, since we do not want noun and adjectives to vary in bias strength. To see this in more detail, let's first consider the female-biased expressions. The mean rating of 2.24 for female-biased adjectives on Scott et al.'s 7-point scale becomes 0.32 (2.24/7) when converted into a proportion, and when converted to a "0=all men, 1=all women" becomes 0.68 (1-0.32). Thus, the mean gender bias of female-biased adjectives is 0.68 (where 1=female), while the mean gender bias of female-biased role nouns in the Misersky et al. norms is 0.718 – i.e. the difference is less than 0.1, as intended. Now, let us turn to the male-biased expressions. The mean rating of 5.28 for male-biased adjec-

tives becomes 0.754 (5.28/7) when converted into a proportion, and when converted to a "0=all men, 1=all women" becomes 0.246 (1-0.754). Thus, the mean gender bias of male-biased adjectives is 0.246 (where 1=female), while the mean gender bias of female-biased role nouns in the Misersky et al. norms is 0.236 – i.e. the difference is again less than 0.1, as intended.

While these calculations do not allow us to make precise comparisons between the bias strength of the sets of role nouns and adjectives used,[1] they do indicate that it is *not* the case that one kind of expression (role noun or adjective) has a much stronger bias than the other. This is as we intended, because we wanted to ensure that male- and female-biased role nouns we tested have gender biases equal in magnitude to those of the male- and female-biased adjectives. This is because our interest lies in comparing how gender bias from role nouns compares that from adjectives, in a situation where the bias strength of nouns and adjectives is otherwise matched.

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 role nouns and adjectives used in Experiments 1 and 2

2.1.3 PROCEDURE AND DATA ANALYSIS The study was implemented using Qualtrics (Provo, UT) and conducted online. Participants, recruited via Prolific, were instructed to fill in the blank in each sentence with one word: they were asked to provide "one word that fills in the blank in a meaningful and natural way." Thus, on target trials, *her*, *his* and *their* are all possible completions. Each sentence was shown on a separate screen. We annotated participants' responses on target trials for whether they filled in the blank with *her*, *his*, *they* or something else.

2.2 RESULTS: EXPERIMENT 1 BY HUMANS. The proportion of time that participants filled in the blank with *her* and *his* in each condition are shown in Figure 1. Although this figure only shows the proportion of trials where participants filled the blank with *her* or *his*, singular *they* was also produced 26.2% of the time, at roughly equal rates across conditions (between 24-28%). Given the roughly comparable *they* rates across conditions, below we focus on the *his/her* data. (For recent work on comprehension of singular *they*, see e.g. Conrod 2022, Arnold, Mayo & Dong 2021, Konnelly & Cowper 2020, Camilliere et al. 2021, Han & Moulton 2022, Gardner & Brown-Schmidt 2024. For a production task on singular *they*, see Kaiser & Post 2025.)

Let's start by looking at the no-adjective conditions which function as a sanity check to see if the role nouns exhibit the predicted biases. Indeed, as can be seen in Figure 1, when only a female-biased role noun is present, participants fill in the blank with *her* over 70% of the time (significantly above chance, intercept-only *glmer* model in R, $p<.001$). When a male-biased role noun is present, participants fill in the blank with *his* over 80% of the time (significantly above chance, $p<.001$). When the role noun is neutral (not biased towards male or female), the rates of *her* vs. *his* completions are at chance ($p>.05$). Thus, these results confirm that our fill-in-blank tasks works, the selected role nouns work as expected, and that participants are paying attention.

Turning to the conditions where the role noun is modified by a female-biased adjective, Figure 2 shows that participants' pronoun choices reflect the gender bias of the *role noun*: A female-biased role noun preceded by a female-biased adjective elicits around 80% *her* completions (sig-

[1] Scott et al. (2019) also normed some professions/role nouns among the large set of words they tested, but they do not include all the role nouns that are in the Misersky et al. (2014) norms.

nificantly above chance, p<.002) and a male-biased role noun preceded by a female-biased adjective still elicits over 70% *his* completions (significantly above chance, p<.01). Neutral role nouns still elicit rates of *his/her* completions that are at chance (p's >0.1), although numerically a female-biased adjective boosts the rate of *her* completions to almost 60%, but this is not significantly above chance (p=0.13). Thus, when a male-biased role noun is modified by a female-biased adjective, participants still tend to assume the referent is male at above-chance rates. This suggests that a role noun's gender bias has more of an effect that the adjective's gender bias.

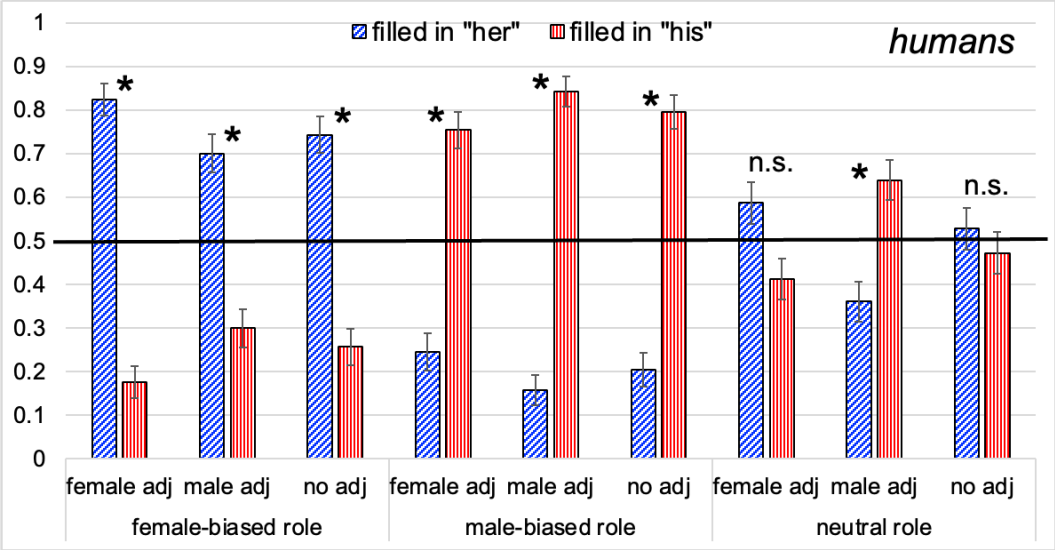

Figure 1. Experiment 1: Proportion of times that participants filled in the blank with *her* vs. *his* (* = significantly different from chance)

Finally, in the conditions where the role noun is modified by a male-biased adjective, we observe something similar: A female-biased role noun preceded by a male-biased adjective elicits around 70% *her* completions (significantly above chance, p<.02) and a male-biased role noun preceded by a male-biased adjective elicits over 80% *his* completions (significantly above chance, p<.001) – in other words, here too the gender bias of the role noun largely determines participants' completions. However, in the neutral role noun conditions, presence of a male-biased adjective is now enough to trigger a higher-than chance rate of *his* completions (p<.02).

In sum, with both female- and male-biased role nouns, regardless of adjective, people tend to produce mostly pronouns that match the gender bias of the role noun. Adjective effects only emerge with neutral nouns, and only with male-biased adjectives. Overall, humans prioritize stereotypical gender information carried by nouns.

2.3 DISCUSSION: EXPERIMENT 1 BY HUMANS. In this experiment, our aim was to test what happens when people encounter conflicting vs. matching cues about a referent's gender: if a role noun and an adjective provide divergent cues (e.g. *an elegant mechanic* or *an imposing flight attendant*), what assumptions do people make about the gender of the referent? The results from Experiment 1 suggest that participants have a strong preference to focus on the stereotypical gender information stemming from the role noun: An adjective whose gender cues mismatch those of the role noun does not override or eliminate the noun's gender bias. When it comes to neutral role nouns that lack a clear gender bias, male-biased adjectives trigger a significant male bias, and female-biased adjectives point to a numerical (albeit non-significant) female bias.

Overall, the results from human participants align best with the *Information-type Asymmetry*

Hypothesis, as they indicate that gender stereotype information carried by the role noun has a bigger impact on people's assumptions about referent gender than does gender stereotype information associated with adjectives. The finding that gender biases encoded on nouns have a stronger effect is compatible with prior work showing that nouns induce stronger stereotyping than adjectives (e.g. Carnaghi et al. 2008, see also Markman 1989). As Carnaghi et al. (2008) put it, "nouns, more than adjectives, lead perceivers to draw inferences that go beyond the information given." We return to this idea in the General Discussion section.

3. Experiment 2: Language task by GPT-4o. To see whether the patterns exhibited by humans also occur with large language models, in this second study we essentially repeated the task that humans had done, but now with OpenAI's GPT-4o (June 2024 paid version). Like humans, we instructed GPT-4o to "provide one word that fills in the blank in a meaningful and natural way" and to provide only one word per blank. Thus, as with humans, we did not explicitly say anything about gender, bias or professions in the instructions, because we did not want to trigger explicit awareness of this (see also Dong et al. 2024 on indirect probing of gender biases of LLMs). We used the same targets and fillers as for humans, and generated data for 90 'participants' by having GPT-4o do the task 90 times. We used the browser interface, crucially with memory set to 'off' and temporary chat turned 'on', so that the model did not learn from or get primed by prior rounds of data generation: As with humans (where each participant saw 27 targets and 33 fillers, but only once), our aim was to make each data generation round as independent from the other rounds as possible. Data analysis was the same as in Experiment 1.

Given that large language models have been trained on human-generated data, one might expect them to exhibit the same kind of patterns as humans – in other words, to be guided more by the stereotypical gender biases of the role noun rather than the adjective. Thus, we may find that GPT-4o exhibits an information type asymmetry, like humans, in line with the Information-type Asymmetry Hypothesis.

However, if the asymmetry we saw with humans – the greater impact of nouns over adjectives – is related to aspects of human cognition, as we might assume given the existing work on the power of nouns in eliciting stereotype inferences, this may not replicate with LLMs. In fact, we may find that LLMs are equally sensitive to both nouns and adjectives, in line with the Information-type Symmetry Hypothesis. Or it may be that LLMs pattern in line with the Gender Asymmetry Hypothesis, such that they are highly sensitive to cues signaling one gender (either male or female), regardless of whether those cues are provided by the noun or the adjective.

Before taking a closer look at the results, a few words are in order about why we chose to use GPT-4o, and why we did not use an open-source model like Llama, for example, which can be trained *without* reinforcement learning from human feedback (RLHF), thus allowing for a clean look at the model's abilities without human interference. Our decision to use GPT-4o was motivated by our aim of getting a sense of what kind of gender bias information, if any, the average user encounters when using LLM systems. Today, most consumer-oriented language models use human feedback to guide the model's learning process. Thus, given that we are interested in getting a sense of what kind of output is typically seen by typical users (in particular, how information about gender stereotypes carried by nouns and adjectives shapes this output and the model's behavior), we opted to use a version that is available to the public. This being said, it would also be very informative to conduct additional research comparing different models.

3.1 RESULTS FOR THE LANGUAGE TASK BY GPT-4O. The proportion of time that GPT-4o filled in *her* and *his* in each condition is shown in Figure 2. (GPT-4o produced singular *they* (not plotted

here) only 1.4% of the time on average, less than humans who produced it ca. 26% of the time.)

First, let's consider the no-adjective conditions. Similar to humans, GPT-4o exhibits the predicted role noun biases: When only a female-biased role noun is present, it fills in the blank with *her* over 95% of the time (significantly above chance, intercept-only *glmer* model in R, p<.001). When a male-biased role noun is present, it fills in the blank with *his* over 90% of the time (significantly above chance, p<.001). With neutral role nouns, *her* vs. *his* completion rates are at chance (p's>.3). Thus, GPT-4o exhibits the expected gender associates with the role nouns.

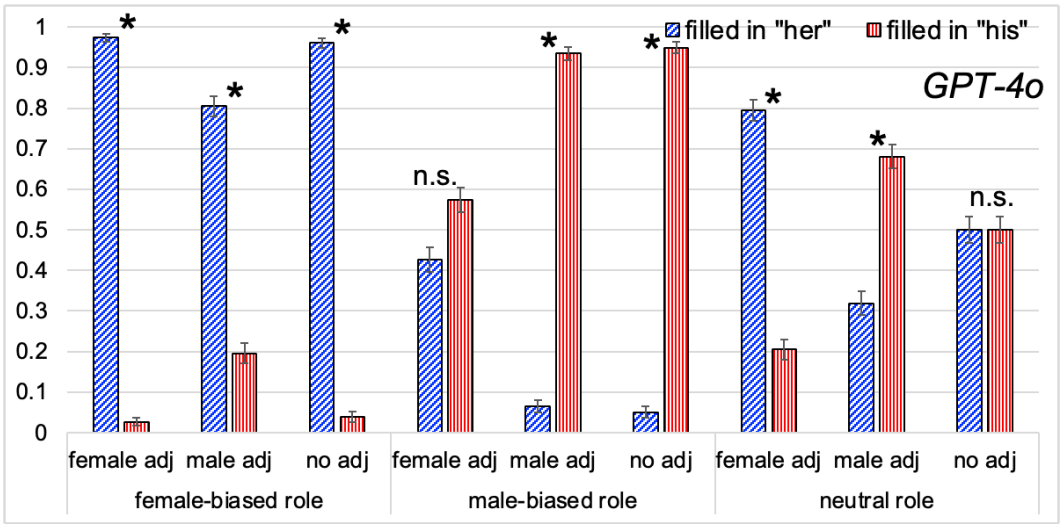

Figure 2. Experiment 2: Proportion of times that GPT-4o filled in the blank with *her* vs. *his* (* = significantly different from chance)

What about conditions where the role noun is modified by a female-biased adjective? Here the data for the cue conflict condition diverges from what we see with humans. A female-biased role noun preceded by a female-biased adjective elicits around 95% *her* completions (significantly above chance, p<.001), similar to humans. However, a male-biased role noun preceded by a female-biased adjective elicits less than 60% *his* completions: although there are numerically more male than female completions (as with humans), with GPT-4o the male preference is not significant (not different from chance, p>0.5). When there is cue conflict due to a female-biased adjective and a male-biased role noun, the female-biased adjective has a detectable effect in that it 'wipes out' the noun's male bias, though its presence is not enough to trigger a female bias.

With a neutral role noun, presence of a female-biased adjective elicits a high rate of *her* completions (80%, above chance, p<.001). Humans showed a pattern in the same direction but it did not reach significance.

Finally, what about conditions where the role noun is modified by a male-biased adjective? Based on the data discussed above for female-biased adjectives, we might expect cue conflict to yield an even split between male and female completions here as well. But instead, the data looks similar to humans: A female-biased role noun preceded by a male-biased adjective elicits around 80% *her* completions (significantly above chance, p<.04) and a male-biased role noun preceded by a male-biased adjective elicits over 90% *his* completions (significantly above chance, p<.001): like with humans, the gender bias of the role noun largely determines the choice of pronoun. Thus, when cue conflict stems from a male-biased adjective and a female-biased role noun, the female bias 'wins out.' In the neutral role noun conditions, presence of a male adjective elicits a higher-than chance rate of *his* completions (almost 70%, p<.02), similar to humans.

3.2 DISCUSSION OF THE LANGUAGE TASK BY GPT-4O. The data from humans and GPT-4o show some unexpected differences. Whereas humans consistently prioritize information from the role noun, it seems that, at least in some contexts, GPT-4o attends more to information from *both* the role noun and the adjective. The differences between humans and GPT-4o become especially clear when we look at conditions where the adjective and noun conflict in their biases. When a female-biased adjective modifies a male-biased role noun, humans still prefer male completions, but with GPT-4o the rate of female and male completions does not differ from chance.

When a male-biased adjective modifies a female-biased noun, humans prefer female completions, as expected since the role noun is female-biased. Perhaps surprisingly, GPT-4o also prefers female completions in this configuration. Why do we see a preference here, in contrast to the situation where a female-biased adjective modifies a male-biased role noun? The only difference is whether the female cue is on the adjective or the role noun: perhaps, echoing the noun bias in humans, GPT-4o gives even more weight to female-biasing cues when they are on the *role noun*. Although further work is needed, the data from GPT-4o provides some initial hints in support of the Gender Asymmetry Hypothesis – specifically, that female-biasing cues receive more weight – combined with signs of the noun bias that humans also exhibit.

As a whole, we find that both humans and ChatGPT are susceptible to gender bias, but humans prioritize gender cues from role nouns over adjectives, while GPT4o seems more relatively sensitive to adjectival information as well, specifically female-biased adjectives.

4. Experiment 3: Image generation by Dall-E 3. Having compared humans and GPT-4o in a language-based task, we also wanted to see how they compare to text-to-image models, such as OpenAI's DALL-E 3. It's worth noting that, under the hood, text-to-image generation is different in various ways from text generation (see e.g. Li et al. 2025a), so *a priori* there's no reason to expect DALL-E 3 to pattern the same way as GPT-4o.

4.1 IMAGE GENERATION DESIGN AND TASK. To assess how gender stereotype information from role nouns and adjectives influences image generation, we had to adjust the fill-in-the-blank task as it is not relevant for an image generation system. Instead, we used prompts like those in (6) to get DALL-E 3 to create images. At the time of data collection, the system default was to provide two images for each item, but there was some variation in this. To ensure consistency of output, we explicitly requested two images each time, as shown by the example prompts in (3).

(6) a. two images: the gymnast [plain role noun]
b. two images: the tough gymnast [attributive adjective + role noun]
c. two images: the gymnast is tough [role noun + predicative adjective]

We intentionally used terse sentence fragments as shown in (6) to minimize potential effects of other information such as thematic role. In addition to 'plain' role nouns without adjectives (e.g. 6a), we tested adjectives in prenominal (6b) and attributive position (6c), but as these two variants did not yield clear differences, we collapse them in the subsequent discussion.

As in Experiments 1 and 2, we manipulated the stereotypical gender properties of role nouns and adjectives. In Experiment 3, we used 20 male-biased role nouns (mean 0.239, SD 0.041, Misersky et al. norms, "0=all men, 1=all women") and 20 female-biased role nouns (mean 0.737, SD 0.048) as well as 20 female-biased adjectives (mean 2.066, SD 0.347, Scott et al. norms, *very feminine* (1) to *very masculine* (7)) and 20 male-biased adjectives (mean 5.44, SD 0.368). In Experiment 3, we used 20 nouns and adjectives, not 27, to keep the number of images manageable (given the larger number of configurations tested here). In addition, we are conducting follow-up

studies (not reported here) using the same-up (nouns and adjectives) as Experiments 1 and 2.

	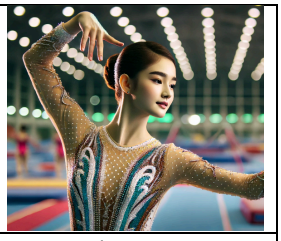
Tough gymnast	Lovely gymnast
	
Emotional nanny	Powerful nanny

Table 2. Examples of images generated by DALL-E 3 and the prompts that generated the images

Instead of using neutral professions as in Experiments 1 and 2, in Experiment 3 used the noun 'person' in the neutral role-noun condition, as the word 'person' is definitionally unmarked for gender. We did not use the noun 'person' in Experiments 1 and 2, because it was judged to sound unusual/awkward in the fill-in-the-blank sentences we used in those experiments. Unlike Experiments 1 and 2, this study did not include any filler items.

Using DALL-E 3 (May/June 2024 paid version), we generated 600 images. Some examples are provided in Table 2. We then coded each image for whether person shown in the image looks more (stereo)typically female or male or if this was unclear. We acknowledge that this is an overly simplistic annotation procedure; in future work we plan to ask participants to rate the resulting images along gradient scales to tap into more fine-grained aspects of the images.

4.2 RESULTS AND DISCUSSION FOR IMAGE GENERATION. The results for the image generation task are shown in Figure 3. The figure shows proportions of male vs. female persons generated, for ease of comparison with Experiments 1 and 2. However, due to the nature of the data, statistical analyses were now conducted on count data using one-way chi-squared tests.

First, let's consider the no-adjective conditions. Similar to Experiments 1 and 2, female- and male-biased role nouns without adjectives exhibit significant female and male biases respectively (over 70% female and 100% male respectively). However, when asked to make an image of a 'person', DALL-E 3 has a very strong *male* default and generated over 90% male images, despite the noun 'person' being gender-neutral. (Naik & Nushi (2023) found that an earlier version, DALL-E 2, similarly generated around 70% male images when prompted with 'person').

Now, let's consider what happens when the role noun is described by a female-biased adjective. A female-biased role noun preceded by a female-biased adjective elicits over 90% female images (significantly more female, one-way chi-squared, $p<.001$). But in a cue-conflict situation, when a male-biased role noun is preceded by a female-biased adjective, we still see over 70% male images (significantly more male, $p<.001$). Thus, the gender of the role noun still has a significant effect, echoing Experiment 1 with humans. With a neutral role noun, presence of a female-biased adjective eliminates the male bias we observed in the no-adjective condition, as

now the images are split between male and female (no significant difference, p>0.5).

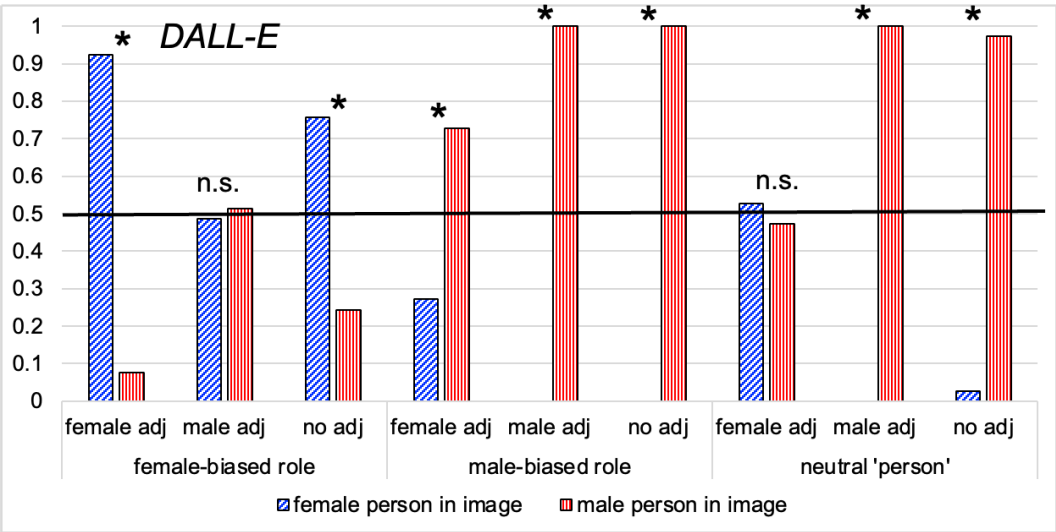

Figure 3. Experiment 3: Proportion of male and female characters generated by DALL-E 3 (* = distribution of male vs. female images differs significantly from an even split)

Finally, let's look at what happens when the role noun is described by a male-biased adjective. When a male-biased role noun is modified by a male-biased adjective, all images are male; the neutral noun 'person' modified by a male-biased adjective also elicits almost all male images. But when a female-biased role noun is modified by a male-biased adjective (cue conflict), the images are now split more evenly between male and female (no significant difference, p> .7).

Thus, with DALL-E 3, the pattern in cue-conflict cases is the 'inverse' of GPT-4o: GPT-4o yielded (i) a *female* bias with a male-biased adjective combined with a *female*-biased role noun and (ii) no significant preference for male vs. female responses when a female-biased adjective was combined with a male-biased role noun – i.e., GPT-4o prioritizes the role noun cue over the adjective cue more when the role noun is female. In contrast, DALL-E 3 yields (i) an equal number of male and female completions when a male-biased adjective was combined with a female-biased role noun and (ii) a *male* bias with a female-biased adjective combined with a *male*-biased role noun. Thus, DALL-E 3 prioritizes the role noun cue over the adjective cue specifically when the role noun is male. Humans, in contrast, consistently prioritize cues from the role noun.

Overall, the data from DALL-E 3 shows that information from the nominal (role nouns) and adjectival domain both contribute, but the asymmetrical contributions differ from GPT-4o.

5. General discussion. The three experiments reported here explore how stereotypical information about gender associated with adjectives and role nouns influences assumptions about referent gender. We tested how people and AI models interpret descriptions of people in situations where there the stereotypical gender associations of (i) role nouns describing professions (e.g. *mechanic, nurse*) and (ii) adjectives describing human properties (e.g. *brave, dangerous, sentimental, graceful*) either align or conflict.

The results from a sentence-completion task show that humans prioritize information from role nouns: When the gender cues of role nouns and adjectives conflict (e.g. *gallant librarian, lovely hunter*), humans tend to interpret the referent in line with the stereotypical gender associations of the role noun (e.g. *librarian* as female, *hunter* as male), as revealed by people's pronoun use. This fits with the Information-type Asymmetry Hypothesis, according to which information

from role nouns vs. adjectives differs in how much it guides people's gender assumptions.

However, the large language model GPT-4o output diverges from human data in showing more sensitivity to adjectival information: While GPT-4o also prioritizes information from the role noun when a male-biased adjective modifies a female-biased role noun, the pattern changes when a female-biased adjective modifies a male-biased role noun: here GPT-4o is at chance and the rate of female and male completions does not differ. These results suggest that while GPT-4o exhibits aspects of the same nominal bias that humans exhibit, it also shows initial hints in favor of the Gender Asymmetry Hypothesis, suggesting that female-biasing cues receive more weight.

Interestingly, in an image-generation task, the T2I model DALL-E 3 shows the 'inverse' behavior in the cue conflict cases compared to GPT-4o: DALL-E 3 prioritizes information from the role noun when a female-biased adjective modifies a *male*-biased role noun, but when a male-biased adjective modifies a *female*-biased role noun, DALL-E 3 produces female and male images equally often. In sum, in cue conflict cases, GPT-4o prioritizes the role noun cue over the adjective when the role noun is female, DALL-E 3 does so when the role noun is male, but humans consistently prioritize cues from the role noun regardless of its gender bias.

The finding that humans prioritize information on role nouns over adjectives brings up the question of *why*. Although our experiments were not designed to address this question directly (and, as mentioned above, in our studies information about professions is correlated with noun status and information about traits/properties is correlated with adjective status), our findings from Experiment 1 align well with existing work suggesting that the noun/adjective distinction plays a key role in guiding how we draw inferences and make generalizations. Prior work suggests that information expressed by nouns tends to be regarded as being more permanent and more important than information expressed by adjectives. For example, Wierzbicka (1986) has suggested that "human characteristics tend to be designated by nouns rather than adjectives if they are seen as permanent and/or conspicuous and/or important" (Wierzbicka 1986, p. 357).

Relatedly, some have argued that nouns tend to favor essentialist thinking and stereotypical inferences more than adjectives. For example, as noted by Ritchie (2021), research in cognitive psychology suggests that "When a noun rather than an adjective is used, both children and adults draw more robust inferences and judge features to be more inheritable, persistent, and explanatory" (Ritchie 2021, p.471). Evidence for this comes from work by Gelman & Markman (1986), Markman (1989), Markman & Smith (cited by Markman 1989), Carnaghi et al. (2008) and others. Although this prior work did not focus specifically on role nouns or male/female gender stereotypes, its findings are compatible with the patterns we observe with humans in Experiment 1. For example, Carnaghi et al. (2008) found that nouns elicit stronger essentializing inferences than adjectives in sentences like *Mark is athletic/Mark is an athlete*: when presented with statements congruent with the noun's/adjective's meaning (e.g. *he runs three times a week*) and asked to rate how strong, stable and resilient the subject's preference is for the activity, the ratings were significantly higher in with noun labels than with adjective labels. Based on a series of studies, Carnaghi et al. conclude that "nouns have a greater likelihood than adjectives to induce stereotype-congruent expectancies" (Carnaghi et al. 2008, p.846).

Although these prior studies do not focus on gender, the finding that noun labels elicit stronger stereotypical inferences than adjective labels fits well with our finding that humans rely more on the role noun than on the adjective when making inferences about the referent's gender, in line with the Information-type Asymmetry Hypothesis. While many questions remain open and more work is needed to test these ideas, the present work provides new insights into how stereotypical gender information from role nouns and adjectives guides the assumptions that hu-

mans and generative AI make about referent gender: While humans, GPT-4o and DALL-E 3 all use stereotypical gender to make inferences, only humans show a consistent pattern of prioritizing nominal information. Although GPT-4o and DALL-E 3 diverge from humans (and each other), their outputs also reveal that whether nominal information is prioritized for making gender inferences depends in systematic ways on the gender of the role noun itself.

References

Arnold, Jennifer E., Heather Mayo & Lisa Dong. 2021. My pronouns are they/them: Talking about pronouns changes how pronouns are understood. *Psychonomic Bulletin & Review* 28, 1688-1697.

Bianchi, Federico, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou & Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* 1493-1504.

Camilliere, Sadie, Amanda Izes, Olivia Leventhal & Daniel Grodner. 2021. They is changing: Pragmatic and grammatical factors that license singular they. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43.

Carnaghi, Andrea, Anne Maass, Sara Gresta, Mauro Bianchi, Mara Cadinu & Luciano Arcuri. 2008. Nomina sunt omina: on the inductive potential of nouns and adjectives in person perception. *Journal of personality and social psychology* 94(5). 839-859.

Conrod, Kirby. 2022. Abolishing gender on D. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 67(3). 216-241.

Dong, Xiangjue, Yibo Wang, Philip S. Yu & James Caverlee. 2024. Disclosure and mitigation of gender bias in LLMs. *arXiv:2402.11190*

Doshi, Rushabh H., Simar S. Bajaj & Harlan M. Krumholz. 2023. ChatGPT: temptations of progress. *The American Journal of Bioethics* 23(4).6-8

Fraser, Kathleen C., Svetlana Kiritchenko & Isar Nejadgholi. 2023. A friendly face: Do text-to-image systems rely on stereotypes when the input is under-specified? *arXiv:2302.07159*.

Friedrich, Felix, Katharina Hämmerl, Patrick Schramowski, Manuel Brack, Jindrich Libovicky, Kristian Kersting & Alexander Fraser. 2024. Multilingual text-to-image generation magnifies gender stereotypes and prompt engineering may not help you. *arXiv:2401.16092*

Gardner, Bethany & Sarah Brown-Schmidt. 2024. Biased inferences about gender from names. *Glossa Psycholinguistics*, 3(1).

Gelman, Susan A. & Ellen M. Markman, Ellen. 1986. Categories and induction in young children. *Cognition* 23. 183-209

Girrbach, Leander, Stephan Alaniz, Genevieve Smith & Zeynep Akata. 2025. A Large Scale Analysis of Gender Biases in Text-to-Image Generative Models. *arXiv:2503.23398*

Han, Chung-hye & Keir Moulton. 2022. Processing bound-variable singular they. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 67(3). 267-301.

Kaiser, Elsi & Claire Benét Post. 2025. Names and Pronouns with and without Gender Features: A Production Study of Singular *they*. *Proceedings of WCCFL 41.* 270-277.

Kaiser, Elsi, & Jamie Herron Lee. 2018. Predicates of personal taste and multidimensional adjectives: An experimental investigation. *Proceedings of WCCFL 35*. 224-231.

Kennedy, Christopher. 2013. Two Sources of Subjectivity: Qualitative Assessment and Dimensional Uncertainty. *Inquiry* 56 (2-3). 258-277.

Konnelly, Lex & Elizabeth Cowper. 2020. Gender diversity and morphosyntax: An account of

singular they. *Glossa: a journal of general linguistics* 5(1). 40.
Kotek, Hadas, Rikker Dockum & David Sun. 2023. Gender bias and stereotypes in large language models. *Proceedings of the ACM collective intelligence conference*. 12-24.
Lasersohn, Peter. 2005. Context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy* 28(6). 643-86.
Li, Jun, Chenyang Zhang, Wei Zhu, and Yawei Ren. 2025a. A comprehensive survey of image generation models based on deep learning. *Annals of Data Science* 12. 141-170.
Li, Jia, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang & Di Wang. 2025b. Fair text-to-image diffusion via fair mapping. *Proceedings of the AAAI Conference on Artificial Intelligence* 39(25). 26256-26264.
Luo, Hanjun, Ziye Deng, Ruizhe Chen & Zuozhu Liu. 2024. Faintbench: A holistic and precise benchmark for bias evaluation in text-to-image models. *arXiv:2405.17814*.
Marin, Alondra & Markus Eger. 2024. Towards Evaluating Profession-based Gender Bias in ChatGPT and its Impact on Narrative Generation. *Proceedings of the AIIDE Workshop on Intelligent Narrative Technologies*.
Markman, Ellen M. 1989. *Categorization and naming in children*. Cambridge, MA: MIT Press.
McNally, Louise and Isidora Stojanovic. 2017. Aesthetic adjectives. In James Young (ed.), *The Semantics of Aesthetic Judgments*, 17-37. Oxford: Oxford University Press.
Misersky, Julia, Pascal M. Gygax, Paolo Canal, Ute Gabriel, Alan Garnham, Friederike Braun, Tania Chiarini, Kjellrun Englund, Adriana Hanulikova, Anton Öttl, Jana Valdrova, Lisa Von Stockhausen & Sabine Sczesny. 2014. Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior Research Methods* 46. 841-871.
Naik, Ranjita & Besmira Nushi. 2023. Social biases through the text-to-image generation lens. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 786-808.
Ritchie, Katherine. 2021. Essentializing language and the prospects for ameliorative projects. *Ethics*, 131. 3460-488.
Sassoon, Galit W. 2013. A typology of multidimensional adjectives. *Journal of Semantics* 30. 335-380
Scott, Graham, Anne Keitel, Marc Becirspahic, Bo Yao & Sara C. Sereno. 2019. The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods* 51: 1258-1270.
Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, & Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems* 33. 12388-12401.
Wan, Yixin, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis & Kai-Wei Chang. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv:2404.01030*
Wan, Yixin, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang & Nanyun Peng. 2023. Kelly is a Warm Person, Joseph is a Role Model: Gender Biases in LLM-Generated Reference Letters. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3730-3748. Association for Computational Linguistics.
Wierzbicka, Anna. 1986. What's in a noun? (Or: How do nouns differ in meaning from adjectives?). *Studies in Language*. 10. 353-389
Zhao, Jinman, Yitian Ding, Chen Jia, Yining Wang & Zifan Qian. 2024. Gender bias in large language models across multiple languages. *arXiv:2403.00277*