What social attitudes about gender does BERT encode? Leveraging insights from psycholinguistics

Julia Watson[1] **Barend Beekhuizen**[2] **Suzanne Stevenson**[1]

[1]Department of Computer Science
University of Toronto
{jwatson, suzanne}@cs.toronto.edu

[2]Department of Language Studies
University of Toronto, Mississauga
barend.beekhuizen@utoronto.ca

Abstract

Much research has sought to evaluate the degree to which large language models reflect social biases. We complement such work with an approach to elucidating the connections between language model predictions and people's social attitudes. We show how word preferences in a large language model reflect social attitudes about gender, using two datasets from human experiments that found differences in gendered or gender neutral word choices by participants with differing views on gender (progressive, moderate, or conservative). We find that the language model BERT takes into account factors that shape human lexical choice of such language, but may not weigh those factors in the same way people do. Moreover, we show that BERT's predictions most resemble responses from participants with moderate to conservative views on gender. Such findings illuminate how a language model: (1) may differ from people in how it deploys words that signal gender, and (2) may prioritize some social attitudes over others.

1 Introduction

Language choices are revealing about speakers' social attitudes – their (evaluative) beliefs, views, and expectations about social phenomena. If a café advertises "gingerbread people," instead of "gingerbread men" (example adapted from Papineau et al., 2022), people may make inferences about the social views of the café owners based on their avoidance of the traditional masculine term. Social attitudes typically surface in less "pointed" but higher stakes scenarios, such as a speaker using the pronoun *they* to refer to a colleague who identifies as nonbinary, reflecting the speaker's acceptance of nonbinary identities.

Much work on the social knowledge encoded in language technology has focused on evaluating whether models encode stereotypical/harmful associations (e.g., Caliskan et al., 2017; Rudinger et al., 2018), and if so, removing them to "de-bias" NLP (e.g., Bolukbasi et al., 2016; Zhao et al., 2018). However, social knowledge permeates language (e.g., Nguyen et al., 2021), and what counts as harmful depends on one's perspective (e.g., Blodgett et al., 2020). To deal effectively with potentially harmful associations in NLP, we need a clear understanding of how social attitudes are linked to the language choices people make, so that we can assess the language choices of our technologies.

Here we seek to understand **what social attitudes a large language model encodes**, specifically social attitudes about gender. To address this question, we draw on datasets from two psycholinguistics studies, both of which included language tasks involving gendered and gender neutral language choices, *and* surveys eliciting the same participants' social attitudes on gender. By explicitly linking people's language choices with their social attitudes, this data enables us to evaluate how social attitudes are reflected in the language choices encoded in an NLP model, and to quantify the extent to which a language model propagates certain views over others (cf. Bender et al., 2021).

In the first study we draw on, Papineau et al. (2022) elicited preferences for feminine, masculine, and gender neutral variants of role nouns, such as *firewoman/fireman/firefighter*, and found that choices to use gendered over gender neutral variants can reflect more rigid views about men's and women's social roles. In the second study, Camilliere et al. (2021) elicited acceptability judgements of singular *they* pronouns in contexts like *My friend$_i$ said they$_i$ would be coming late to dinner*. They found that lower acceptability ratings of singular *they* are associated with less acceptance of nonbinary people. It is important to determine if language models make similar choices to these, since if they do, they may spread and reinforce such attitudes, which may contribute to gender stereotyping (Sczesny et al., 2016), or nonbinary erasure

(Cao and Daumé III, 2020; Dev et al., 2021).

We use the datasets from these two experiments to evaluate the large language model BERT (specifically, BERT-base-uncased, Devlin et al., 2019). We focused on a masked language model because such models can readily mimic the linguistic tasks in these experiments. We selected BERT specifically because it has been widely deployed and thoroughly evaluated in the computational linguistics literature, which facilitates comparison with past studies. Additionally, our focus on the light-weight BERT-base-uncased allowed for more experimentation, letting us carefully evaluate numerous experimental conditions across multiple participant groups. Although we focus on masked language modeling and BERT,[1] our approach for relating linguistic behaviour to social attitudes is generalizable, and can readily be extended to other models or tasks.

For each of the datasets we consider, we explore the following two research questions:

RQ1: Is BERT influenced by the same linguistic cues as people in language choices that signal gender?

We address RQ1 by studying whether BERT takes into account the linguistic cues shown in these psycholinguistic experiments to influence people's word choices, generating language involving gender consistently with human expectations. We examine pragmatic factors that have not been previously explored in the use of gendered and gender neutral language by large language models.

RQ2: What social attitudes about gender are reflected in BERT's word preferences?

In exploring RQ2, we consider BERT's preferences compared to those of participants grouped by their social attitudes, as revealed in the survey data. In doing so, we undertake the first analysis of BERT's word preferences in gender-relevant language that reveals the social attitudes that BERT's choices are most aligned with.

To preview our results, we find that BERT's behavior reflects factors that shape human lexical choices of gendered and gender neutral language, but may not weigh them in the same way people do. Moreover, BERT's predictions most resemble responses from participants with moderate to conservative views on gender. Such findings illuminate how a language model: (1) may differ from people in how it deploys words that (implicitly or explicitly) signal gender, and (2) may prioritize (and propagate) some social attitudes over others.[2]

2 Related Work

Much research has explored what NLP models have learned about language and gender. Related to our work on role nouns, prior work has shown that word embeddings encode stereotypical gender associations for occupation words like *nurse* and *doctor* (Bolukbasi et al., 2016; Caliskan et al., 2017). Other papers have found evidence of similar associations in coreference resolution, with models performing better on examples like *she* (rather than *he*) co-referring with *nurse* (Rudinger et al., 2018; Zhao et al., 2018). In contrast, we study language model choices between explicitly gendered and gender neutral variants of role nouns, such as *firewoman/fireman/firefighter*, comparing model choices to those of people with differing social attitudes.

A focus of much recent work is the processing of gender neutral pronouns by NLP systems in the context of reference to nonbinary individuals. Research has shown that while coreference systems are sensitive to some of the same cues to acceptability of singular *they* as people are (Baumler and Rudinger, 2022), language models can have difficulties with gender neutral singular pronouns (Dev et al., 2021; Brandl et al., 2022). Cao and Daumé III (2020) found that removing explicit cues to gender (e.g., replacing gendered pronouns with neutral variants) resulted in worse performance on a coreference resolution task (Webster et al., 2018). We extend such work by looking at an additional factor in acceptability of singular *they*, and (as with role nouns) relating language model predictions to social attitudes.

Crucially, although some of the above papers compare NLP behavior to human responses generally (e.g., Caliskan et al., 2017; Brandl et al., 2022), none draw on data, as we do here, that directly links experimental participants' language choices

[1] In the remainder of the paper, for ease of reading we use the term 'BERT' to refer to the particular BERT-base-uncased model.

[2] The code for all analyses is available at https://github.com/juliawatson/bert-social-attitudes. The data for the analyses in Part 3 is available at https://github.com/BranPap/gender_ideology/; the data for the analyses in Part 4 was obtained from the authors (Camilliere et al., 2021).

and social attitudes. Cao and Daumé III (2020) indirectly highlight how model choices reflect social attitudes, by showing poor performance on data written by/about trans people. We make this link more explicit, across both linguistic phenomena we study, by comparing model predictions to linguistic judgements by participants for whom we also have survey data reflecting their social attitudes.

We do this in the context of much work on language and social attitudes. Sociolinguists have studied the subtle yet pervasive ways that language communicates social meaning around gender (e.g., Eckert, 2012; Meyerhoff, 2014), and raised concerns about how this is handled in NLP (Nguyen et al., 2021). Discourse Analysis emphasizes words as social categories (e.g., Stokoe and Attenborough, 2014), which computational work has operationalized to study online attitudes about gender (LaViolette and Hogan, 2019; Li and Mendelsohn, 2019). Past computational work in this vein has studied variation in use of gendered vs. gender neutral terms across online communities (CH-Wang and Jurgens, 2021). Here, we take this sociolinguistic lens to evaluating two different kinds of gendered and gender-neutral language choices in large language models.

3 Gendered/Gender Neutral Role Nouns

We first evaluate BERT using data from a psycholinguistic experiment by Papineau et al. (2022),[3] which found different usage patterns of gendered and gender neutral role nouns, such as *firewoman/fireman/firefighter*, when applied to women's and men's names as referents. This data enables us to address our first research question (RQ1 above) by examining the extent to which BERT deploys role nouns in a manner consistent with human usages given the linguistic cue of gendered names. While much work has looked at learned gendered associations with role nouns in language models, we know of no work that assesses model choices among gendered and gender neutral variants compared to human preferences.

Papineau et al. (2022) also solicited each participant's responses to a questionnaire on gender and social roles. The questionnaire data enables us to address RQ2 by probing whether BERT's behavior aligns more with participants having conservative, moderate, or progressive social attitudes on gender.

[3] https://github.com/BranPap/gender_ideology/

3.1 Psycholinguistic data on role nouns

Papineau et al. (2022) used a forced-choice production task in which 301 participants (L1 English speakers in the US) were asked to pick the most appropriate variant of a role noun set for sentences of the form "NAME is a ____ from STATE"; e.g.:

3-way split:
`Sally is a (firewoman, fireman, firefighter) from Utah.`

2-way split:
`David is an (actor, actress) from Kansas.`

The relevant difference in the critical stimuli was that the subject was either a common woman's name or a common man's name, and the experiment aimed to see how the gender of the name affected people's choice of role noun variant. (Details on the names and how they were selected can be found in Appendix A.1).

The stimuli included 20 different sets of role nouns: 14 have a **3-way split** between feminine [FEM], masculine [MASC], and gender neutral [G-NEUT] variants, and 6 have a **2-way split** between a FEM variant and a variant that can be MASC and/or G-NEUT. (Appendix A.1 lists all the role noun sets.) Because of this difference, we analyze the 3-way and 2-way role noun sets separately.

Papineau et al. (2022) also scored each participant given their responses on the Social Roles Questionnaire of Baber and Tucker (2006), in which higher scores mean more rigid views about the social roles of men and women. Following Papineau et al. (2022), we refer to participants with higher scores (more rigid views) as having more conservative attitudes about gender. For our analyses, we grouped participants into three bins based on this score: those with progressive gender attitudes (lowest third of scores; n=90), moderate gender attitudes (middle third; n=90), and conservative gender attitudes (highest third; n=91).[4] Appendix A.2 provides details on this survey, and how we grouped participants based on their responses. Figure 1a shows, for each of the participant groups, the average proportion of responses of FEM/MASC/G-NEUT variants for the 3-way role nouns, given a woman's or a man's name.

[4] Papineau et al. (2022) further divided participants by gender attitudes *within* political affiliation, but this yields several very small participant groups.

3.2 Calculating BERT's preferences

To mimic human behavior on the forced-choice fill-in-the-blank task, we compute BERT's relative probability, $P(V|C)$, for each variant V in a role noun set (e.g., *firewoman/fireman/firefighter*) in the context C of a given sentence frame (e.g., "Sally is a _____ from Utah"). Normalizing these so they sum to 1 across the variants of a role noun set yields a value analogous to the proportion of human participant responses for each of the FEM/MASC/G-NEUT variants.

BERT can be used as a masked language model to generate such probabilities; however, the direct method of masking the target – e.g., giving BERT "Sally is a [MASK] from Utah" and comparing its probabilities of *firewoman/fireman/firefighter* for the mask – is not appropriate. Some role noun variants differ in their number of words (e.g., *police officer* vs. *policeman*), and this is compounded by BERT breaking many words into multiple word pieces (e.g., *firefighter* is *fire* plus *##fighter*). This often leads to an unfair comparison of $P(V|C)$ over varying numbers of masked items for V.

To deal with this issue, we apply Bayes rule:

$$P(V|C) = \frac{P(C|V)P(V)}{\sum_V P(C|V)P(V)} \quad (1)$$

where $\sum_V$ is calculated over the variants in a given role noun set (e.g., *firewoman*, *fireman*, *firefighter*). Because the context C – the words in the sentence other than the role noun – has the same words in the case of all variants of a role noun set, $P(C|V)$ can be compared fairly across variants of a set.

We set the prior term $P(V)$ for a role noun variant V (e.g., *firefighter*) equal to its frequency divided by the summed frequencies for all variants for that role noun set (e.g., *firewoman*, *fireman*, *firefighter*). To reflect BERT's exposure to the role nouns, we use frequencies consistent with BERT's training data; see Appendix A.3.

To approximate $P(C|V)$, we adopt the approach from Nangia et al. (2020), which adapts the *pseudo-log-likelihood* scoring method from Salazar et al. (2020).[5] This method calculates the probability of each word c_i in the context C, from the entire sentence frame, including the variant V. Let S be the full sentence, such as *Sally is a firefighter from Utah*, given a variant role noun V (here, *firefighter*). Then we define $P(C|V)$ as:

$$P(C|V) \doteq \prod_{c_i \in C} P(c_i|S \backslash c_i) \quad (2)$$

where $P(c_i|S\backslash c_i)$ is BERT's probability of the context word c_i given the remainder of the sentence. For example, for the context word *Sally* in *Sally is a firefighter from Utah*, we would feed into BERT "[MASK] is a firefighter from Utah", and look at the probability of *Sally* in masked position. We do this for all context words c_i (these do not include the role noun term), and take the product.

While Equation (1) only indirectly predicts the variant role nouns from the context ($P(V|C)$), by predicting the context words from the rest of the sentence including the role noun ($P(C|V)$), it draws on the same learned associations of BERT that we want to tap into – i.e., the associations between a gendered referent (*Sally* or *David*) and the FEM/MASC/G-NEUT variants of a role noun.[6]

The data provided by Papineau et al. (2022) for each stimulus sentence includes the subject name (e.g., *Sally*) and the role noun set (e.g., *firewoman*, *fireman*, *firefighter*), but does not include the state name used (e.g., *Utah*). We average $P(V|C)$ from Equation (1) across 50 versions of each stimulus sentence with each of the 50 US state names.

3.3 Results on role noun selection

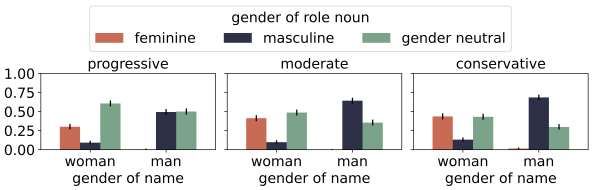
We focus on results for forms with a 3-way split (e.g., *firewoman/fireman/firefighter*) and (for space reasons) summarize the differences found for forms with a 2-way split (e.g., *actor/actress*). (Complete 2-way results are in Appendix A.5.)

For comparison to the human data in Figure 1a, we plot BERT's averaged probabilities in Figure 1b. To assess the degree to which frequency may be driving BERT's predictions, we plot in Figure 1c the predictions from a frequency baseline (using the frequency prior from Equation (1)).

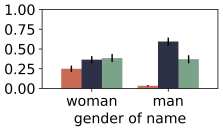
In addition, we compute the average log likelihood, according to BERT's probabilities, of responses of participants in each gender attitudes group – progressive, moderate, and conservative –

[5] Differently from Nangia et al. (2020), we use $P(C|V)$ in the context of calculating the posterior probability $P(V|C)$, which takes into account the prior $P(V)$. Since Nangia et al. (2020) wanted a "score" that was independent of the frequency of the variant, they directly compared values of (their equivalent of) $P(C|V)$ across different variants.

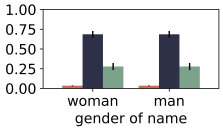
[6] Use of Equation (1) is further justified by its strong, significant correlation ($r = .76$) with the direct method of masking the role noun to calculate $P(V|C)$, for the 8 role noun sets for which the direct method was possible; see Appendix A.4.



(a) Participant responses (Papineau et al., 2022).



(b) BERT predictions



(c) Frequency prior

Figure 1: (a) Participant responses, (b) BERT predictions, and (c) frequency-based predictions, for 3-way role nouns (e.g., *firewoman/fireman/firefighter*) by gender attitudes (progressive/moderate/conservative) and by subject name (woman's/man's).

shown in Table 1.[7] This identifies which participant group's responses are best predicted by BERT.

3.3.1 RQ1 and role nouns

Our first research question asks to what extent BERT is sensitive to the same linguistic cues as people in making language choices involving gender. In the stimulus sentences here, the only cue for selection of a role noun variant is the gender associations with the subject name (e.g., *Sally* or *David*). For RQ1, then, we aim to see if BERT's responses for women's and men's names follow a similar pattern to the general patterns across all gender attitude groups of participants.

The human data in Figure 1a clearly shows that, across all participant groups, responses depend on the (presumed) gender of the subject name. For ease of presentation, we refer to FEM and G-NEUT role noun forms as "congruent" with women's names, and MASC forms as not congruent, and the reverse for men's names (MASC and G-NEUT forms congruent, and FEM not). Across all groups in the human data, for each type of name, there are few forms that are not congruent – a small proportion of MASC forms for women's names (black bars) and almost no FEM forms for men's names (orange bars). For the congruent forms, neither is completely dominant for either women's or men's names – they vary between being about equally represented, to one of the forms being about twice

[7]Total log likelihood is not appropriate because the participant groups differed in the number of observations.

gender attitudes group	loglik for women's names	loglik for men's names	loglik for all data
prog	-1.43	-1.01	-1.23
mod	**-1.33**	**-0.75**	**-1.05**
cons	-1.40	-0.78	-1.08

Table 1: Average log likelihood of data from participants in different gender attitude groups (prog[ressive], mod[erate], cons[ervative]), based on BERT's predictions (3-way role nouns). (Number of observations is 1,260, 1,260, and 1,274, respectively.) Higher scores indicate better fit; best per column (stimulus type) is bolded.

the rate of the other (G-NEUT and FEM for women's names [green and orange bars], MASC and G-NEUT for men's names [black and green bars]).

Figure 1b shows that BERT is also clearly sensitive in its role noun preferences to the gendered associations with the subject names: The patterns are different across men's and women's names, so BERT's behavior is not due solely to frequency of the variant forms. BERT shows a human-like pattern for men's names, with very few non-congruent FEM forms, and neither of the MASC or G-NEUT forms completely dominating. However, BERT does not match human behavior on the women's names. Although congruent forms (FEM and G-NEUT) are close to equally represented, the non-congruent MASC form is highly over-represented in comparison to any participant group. Reflecting this pattern, BERT's predictions have a worse fit for women's names compared to men's names, for all three participant groups (Table 1, columns 2 and 3). The frequency baseline (Figure 1c) suggests this worse performance on women's names may be due to the lower frequency of the congruent FEM and G-NEUT forms as compared to the non-congruent MASC forms. For BERT, unlike for humans, the cue of women's names is not sufficient to overcome the frequency bias towards these MASC forms.

Interestingly, we find a different pattern on the 2-way forms, such as *actor/actress*. (See full results and discussion on 2-way forms in Appendix A.5.) There, BERT has a good fit to general human patterns for women's names, but a worse fit on men's names, again overusing non-congruent forms (in this case, FEM). On close examination, this is due to two specific items – *heiress* and *hostess* – being inappropriately preferred for men's names, likely due to specific word co-occurrence patterns; e.g.,

the congruent form *heir* is typically used in a modified context (such as *heir to X*), and not as a bare noun (as in the stimuli here).

Our conclusion on RQ1 is that while BERT, like humans, seems to use gender associations with names to help guide selection of FEM/MASC/G-NEUT role nouns, BERT does not weigh cues to referent gender in the same way that people do. BERT appears to be more influenced by form frequency and other low-level contextual information, such that non-congruent forms (MASC for women's names, FEM for men's names) may be overused compared to humans. This means that BERT is at risk of using gendered and gender neutral noun variants inconsistently with human expectations.

3.3.2 RQ2 and role nouns

Our RQ2 asks which gender attitudes group BERT's predictions most resemble. For 3-way stimuli containing women's names, BERT predicts somewhat more G-NEUT forms than FEM forms, in line with the moderate gender attitude group (which has the best log likelihood score; Table 1, column 2). BERT performs worst on responses from participants with progressive views, because they have both a much larger proportion of G-NEUT forms, and a smaller number of MASC forms, compared to BERT.

For stimuli containing men's names, BERT predicts high rates of MASC role nouns, with substantially more than G-NEUT forms. This most closely resembles responses from participants with moderate and conservative views on gender, since progressives, by contrast, have roughly equal proportions of MASC and G-NEUT forms for men's names. Supporting this, BERT's predictions had the highest (best) log likelihood on the moderate and conservative groups, with minimal differences between them (Table 1, column 3).

On forms with a 2-way split, there are minimal differences between the participant groups, and BERT performs similarly on each of them.

Overall then for RQ2, across men's and women's names, BERT performs most like participants with moderate and conservative social attitudes on gender roles (Table 1, column 4). For role nouns with a 3-way FEM/MASC/G-NEUT split, this is especially due to its high probability for MASC forms for both women's and men's names. This means that BERT is at risk of conveying (and propagating) rigid social attitudes on gender in its use of role nouns.

4 Acceptability of Singular *they*

Use of singular *they* has been evolving in English, from acceptability only with generic or quantified referents (1), to use with non-gendered referents (2), to antecedents of any gender (especially use with nonbinary referents or those of unknown gender) (3) (Konnelly and Cowper, 2020):[8]

1. **Non-innovative**: only generic or quantified antecedents (e.g., *every dentist*)
2. **Innovative**: those in (1) plus non-gendered antecedents (e.g., *the dentist*, *my friend*)
3. **Super-innovative**: those in (2) plus gendered nouns and names (e.g., *my sister*, *Sophia*)

Moreover, psycholinguistic experiments have found that acceptability of singular *they* in the latter two cases is correlated with various measures of openness and familiarity with gender diversity (Ackerman, 2018; Camilliere et al., 2021).

We evaluate BERT on data from Camilliere et al. (2021),[9] who show that both gender and social closeness of antecedents influence participants' acceptability of singular *they*. Here we address our RQ1 by seeing if BERT's assessment of singular *they* is sensitive to social closeness, a subtle factor that has figured in theories of pronoun use, but has not been shown before in a language model.

Camilliere et al. (2021) also collected data on gender attitudes through surveys of the same participants. For RQ2, we compare BERT's pattern of responses to participant groups of both varying linguistic progressiveness (with respect to the groupings above), and differing social attitudes, to assess who BERT's behavior is most aligned with.

4.1 Psycholinguistic data on *they*

Camilliere et al. (2021) asked 160 participants (L1 English speakers from the US) to judge how naturally *they* referred to different kinds of antecedents, using stimuli such as:

`NP said they would be coming late to dinner.`

where NP was replaced with one of the types of antecedents shown in Table 3. (Note that singular *they* cannot be used with inanimates – **The cup$_i$ fell and they$_i$ broke* – hence the inanimate items

[8] Labels of, and examples from, these three stages are taken from Camilliere et al. (2021), for ease of comparison.

[9] The authors provided us this data upon our request.

are controls.) Including all versions of critical and control trials yields 335 sentences for evaluation of BERT.

Figure 2a shows the results by antecedent type from Camilliere et al. (2021). For their analyses, Camilliere et al. (2021) grouped participants based on their ratings into the stages of singular *they* usage described above. (#Non-innovators=43; #Innovators=89; #Super-innovators=16.)

In addition, Camilliere et al. (2021) had participants complete surveys probing social attitudes on gender. Responses on two of these were predictive of *they* ratings, such that more acceptance of and more familiarity with nonbinary genders were associated with more acceptability of singular *they*.

See Appendix B.1 and Appendix B.2 for more details on the Camilliere et al. (2021) data.

4.2 "Naturalness" of *they* in BERT

We use surprisal, $-\log P(they|context)$, as BERT's assessment of *they* in context. Much work in psycholinguistics shows that surprisal captures human expectations for words in processing sentences (e.g., Hale, 2001; Smith and Levy, 2008), so it works well for comparing BERT to human ratings of naturalness here. We feed into BERT the 335 stimuli from Camilliere et al. (2021), masking *they*, as in:

My friend said [MASK] would be coming late to dinner.

and calculate the surprisal of *they* from its probability in masked position.

While people were asked to rate how naturally the pronoun *refers to the target antecedent*, BERT's probability of *they* may not correspond to that reading. However, the stimuli are biased to such a reading (rather than *they* referring to an antecedent outside the sentence); moreover, our results find that BERT's behavior changes depending on both closeness and gender of the target antecedent, strongly suggesting it takes that linking into account.

In our statistical analyses below (regression and correlations), we directly use the surprisal values, $-\log P(they|context)$. However, surprisal values are awkward for visualization purposes, because *higher* surprisal values from BERT correspond to *lower* naturalness ratings from humans. For ease of comparison to the human ratings in Figure 2a, we graph adjusted surprisal values for BERT in Figure 2b. These are the average surprisal values subtracted from a constant (we used 8 to yield a similar scale to the human data), such that higher adjusted surprisal for BERT corresponds to higher naturalness for humans. The unadjusted surprisal scores are graphed for comparison in Appendix B.3.

	β	Std. Error	p-value
close	0.51	0.07	$p \ll 0.0001$
gendered	0.34	0.07	$p < 0.0001$

Table 2: Linear mixed-effects regression predicting (unadjusted) surprisal from BERT for *they*, as a function of whether the antecedent is socially close or gendered.

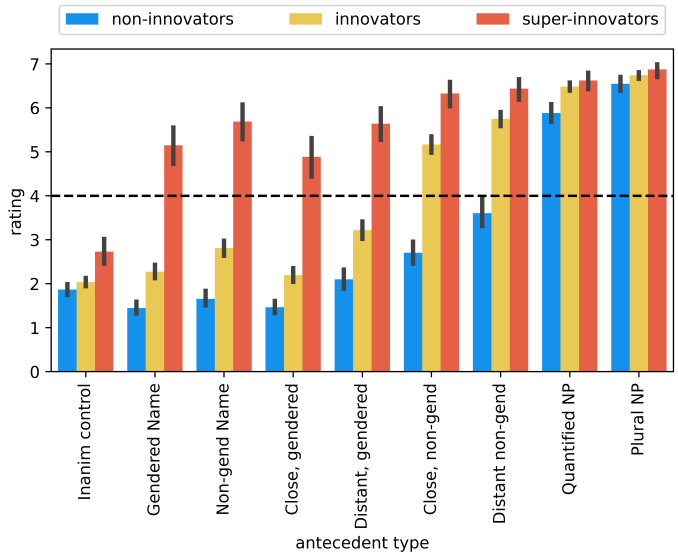
4.3 Results on singular *they*

4.3.1 RQ1 and singular *they*

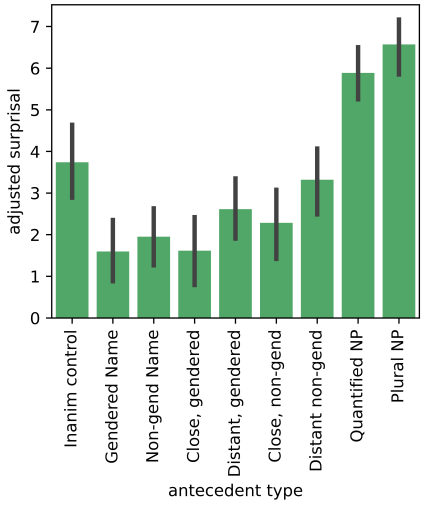
We start by seeing whether BERT is sensitive to the same factors as humans in assessing the acceptability of *they* in context. Comparing Figures 2a and 2b shows that, like people across all groups, BERT rates *they* most highly for both plural antecedents (e.g., *the dentists*) and singular quantified ones (e.g., *every dentist*), showing that in addition to singular/plural, BERT is sensitive to the quantification distinction. However, BERT does not seem as sensitive as people to the cue of (in)animacy, since it did not find *they* in the control items with inanimate subjects (e.g., *the cup*) as bad as people do. (As noted, singular *they* cannot refer to inanimate subjects.)

A benefit of the Camilliere et al. (2021) data is that their experiment manipulated social closeness (e.g., *my friend* vs. *the dentist*) and gender (explicitly gendered or not, e.g., *my friend* vs. *my sister*) of antecedents; the relevant contrasting four conditions are shown in bold in Table 3. Following Camilliere et al. (2021), we conducted a linear mixed effects regression using closeness and gender as fixed categorical factors (with sentence frames as random effects with random intercepts), predicting BERT's surprisal for *they* in the four relevant conditions (160 items). Results in Table 2 show that these factors influence BERT's predictions as they do in humans: *they* is significantly less likely for antecedents that are socially close (vs. distant) or gendered (vs. gender neutral).

For RQ1, we find that BERT is mostly well matched to humans in its basic consideration of the factors influencing naturalness of singular *they*. On one hand, BERT is not as responsive to inanimacy as people are (cf. behavior in the inanimate control condition). On the other hand, BERT – like



(a) Human judgements from Camilliere et al. (2021), by participant group and antecedent type (**higher values mean more natural**).



(b) Adjusted BERT surprisal, by antecedent type (**higher values mean more probable**).

Figure 2: (a) Human judgements and (b) BERT predictions, by antecedent type. Examples of antecedent types are given in Table 3. Error bars in all graphs show 95% confidence intervals. (Figure 4 in Appendix B.3 repeats this figure with unadjusted surprisal values from BERT).

Antecedent Type	Example
Inanimate Control	*The cup*
Gendered Name	*Sophia*
Non-gendered Name	*Taylor*
Close, Gendered	*My sister*
Distant, Gendered	*The actress*
Close, Non-gendered	*My friend*
Distant, Non-gendered	*The dentist*
Quantified NP	*Every dentist*
Plural NP	*The dentists*

Table 3: Types of antecedents in stimuli from Camilliere et al. (2021). Those in bold are used in the analysis of social closeness and gender.

Grouping of participants ...by linguistic stage	r	p-value
non-innovators	–0.62	$p \ll 0.0001$
innovators	–0.57	$p \ll 0.0001$
super-innovators	–0.38	$p \ll 0.0001$
...by gender attitudes	**r**	**p-value**
low nonbinary acceptance	–0.59	$p \ll 0.0001$
med nonbinary acceptance	–0.60	$p \ll 0.0001$
high nonbinary acceptance	–0.43	$p \ll 0.0001$

Table 4: Correlations between (unadjusted) surprisal from BERT and mean rating of each participant group, on 335 stimuli.

humans – is sensitive to number and quantification of antecedents, as well as to gender (as previously considered in NLP, e.g., Baumler and Rudinger, 2022). Moreover, we show that BERT also takes into account the linguistic signal of social closeness – exemplified by the contrast between NPs such as *my friend* vs. *the dentist* – a subtle factor not previously demonstrated before.

4.3.2 RQ2 and singular *they*

Again, our second research question asks which group of human participants – in terms of social attitudes – BERT's predictions most resemble. For each participant group identified in Camilliere et al. (2021) – non-innovators, innovators, and super-innovators – we compute the average rating for each of the 335 stimuli, and then take the Pearson correlation between these ratings and BERT's surprisal for *they* in each stimulus; see the top panel of Table 4.[10] While BERT's predictions significantly correlate with human judgements for all linguistic stages of participants, the correlation is strongest for non-innovators (-0.62), and much weaker for super-innovators (-0.38).

This effect is supported by a visual comparison of the pattern of results shown in Figures 2a

[10] Correlations use all 9 conditions of Figure 2; the same pattern holds with inanimate controls excluded, as well as when using raw probability in lieu of surprisal (see Appendix B.4).

and 2b. Like the non-innovative group (Figure 2a, blue left bars), BERT predicted *they* as much more acceptable for plural and quantified antecedents compared to (close and distant) non-gendered antecedents. In contrast, the super-innovative group (red right bars) gives similar ratings for these four antecedent types (error bars are overlapping). The innovators (yellow middle bars) are in-between, but closer to non-innovators.

While these stages of singular *they* usage are known to reflect social attitudes about gender (Bjorkman, 2017; Konnelly and Cowper, 2020), we wanted to inspect the extent of this connection. We calculated the nonbinary acceptance and familiarity scores of each participant group, since Camilliere et al. (2021) found these factors were overall predictive of naturalness of singular *they*. We found that only the super-innovative group differs significantly in its scores from the other two groups. (See Appendix B.5 for details.)

For a more direct connection to gender attitudes, we split all participants into our own 3-way grouping, based on a low (n=42), medium (n=80), or high (n=24) nonbinary acceptance score.[11] (Details on the nonbinary acceptance survey, and how participant scores were used to group participants, can be found in Appendix B.2.) We repeat the correlation of BERT surprisal for *they* with those of each of these nonbinary-acceptance participant groups; see Table 4. Here, with groupings based explicitly on social attitudes about gender, we find that BERT's behavior is least similar to those most accepting of nonbinary individuals.

Overall then for RQ2, BERT's predictions most resemble those of the non-innovative group of participants in the identified stages of singular *they* acceptance, and least resemble those of super-innovators and those who are more accepting of nonbinary individuals. Again, as with gendered and gender neutral role nouns, we find that BERT's learned knowledge of gender neutral pronoun usage may encode harmful and exclusionary attitudes.

5 Conclusions

In this project, we develop an approach for evaluating the social attitudes encoded in large language models. To do this, we leverage experimental data from psycholinguistics, and compare the predictions of a language model to responses from participants with different social attitudes. This contrasts with much past work on bias in NLP, which has often tested whether models encode stereotypical associations or not, rather than taking a comparative approach to learned associations, and considering how those may relate to social attitudes.

Moreover, we applied our approach to two psycholinguistics datasets, on very different linguistic phenomena involving gender, and obtained very similar results on both. We found that BERT's predictions for role nouns (e.g., *firewoman/fireman/firefighter*) most resemble responses from participants with moderate to conservative views about the social roles of women and men. For singular *they*, we found that BERT's predictions most resemble acceptability judgements from participants with low to moderate nonbinary acceptance scores. Thus when deployed, BERT may propagate language choices that convey such attitudes, reifying them as standard (Blodgett et al., 2020), and reinforcing existing gender inequities (Bender et al., 2021).

Additionally, our analyses have expanded the kinds of linguistic phenomena studied in large language models. Past work has not assessed model preferences for gendered vs. gender neutral variants for role nouns, a lexically rich domain (e.g., Stokoe and Attenborough, 2014). We find that BERT, like people, is sensitive to gender associations in deploying role noun variants, but does not always do so consistently with human expectations. Because misgendering by NLP technology is known to cause harm (e.g., Dev et al., 2021), role nouns constitute an important domain for future study. Our study of singular *they* also extended existing research. Building on past work on singular *they* in NLP (Cao and Daumé III, 2020; Dev et al., 2021; Baumler and Rudinger, 2022; Brandl et al., 2022), we examine how probable language models find *they* when referring to socially close (e.g., *my friend*) vs. socially distant (e.g., *the dentist*) antecedents. We find that, like humans, BERT is sensitive to this contrast, contributing to a growing body of research on the social and pragmatic knowledge learned by large language models.

6 Limitations

In this paper, we developed an approach for evaluating how large language models encode social attitudes about gender, and we applied that approach to evaluate BERT-base-uncased. Because the goal

[11] We chose this measure rather than nonbinary familiarity because we think it is more reflective of social attitudes.

of this paper was ethical in nature, limitations on the generalizability of our approach and findings entail ethical risks. With this in mind, we discuss both limitations and risks in this section. We first discuss limitations related to data, and then discuss those related to models and tasks. For both data and models/tasks, we consider general limitations of our approach, as well as more specific limitations of how we applied the approach here.

6.1 Limitations related to data

Just as it is not possible to create a single benchmark for all language understanding (Raji et al., 2021), it is not possible to create a single, definitive dataset that relates language choices to social attitudes. Human experimental data is always limited by practical considerations and cannot test every condition of theoretical interest; e.g., in the role nouns dataset, there were no conditions with gender neutral names, while in the singular *they* dataset, there was no comparison to neopronouns (e.g., *xe/xem*). Additionally, because past work has found that model preferences may vary across similar linguistic contexts (Delobelle et al., 2022), it may be the case that BERT's predictions would correlate differently with human responses on other variations on the stimuli. Relating model preferences to human behaviour will always be limited by the amount of human data that can be obtained.

Moreover, datasets are always situated in a perspective, emphasizing some people or views over others (e.g., Barrowman, 2018; Chasalow and Levy, 2021). For example, both datasets we consider focus on first language English speakers from the United States, and the specific relationship between social attitudes and linguistic choices captured by those datasets may not generalize outside that context. Languages other than English may have extensive grammatical gender systems, or classification systems that include social roles, among other linguistic devices, which interact to yield rich mechanisms for expressing social attitudes around gender. Even within English speakers in the US, how language signals social attitudes about gender may vary across groups and social contexts. (In fact, Papineau et al. (2022) found that Republicans with progressive social attitudes about gender did not use more gender neutral forms the way Democrats did; other, more fine-grained differences likely also exist.)

Additionally, relating social attitudes about gender to linguistic choices requires some method for measuring social attitudes. Since conceptions of gender are so diverse and culturally variable, no single measurement would be appropriate for all contexts. For example, in one of the datasets we used, a survey for measuring social attitudes about gender asks participants to evaluate statements about stereotypical social roles of men and women, which are likely culturally specific (e.g., "A father's major responsibility is to provide financially for his children") (Baber and Tucker, 2006).

In evaluating language technology, a focus on associations between linguistic choices and social attitudes limited to particular linguistic and cultural contexts risks prioritizing the social knowledge from those communities, and imposing that in other communities when language technology is deployed. To support the creation of inclusive technology, the research community will need to prioritize generation of datasets like the two we drew on here – i.e., ones explicitly connecting linguistic choices to social views – across more languages and cultural contexts.

6.2 Limitations related to models and tasks

There are also several limitations related to the models and tasks considered. First, we evaluated only one model (BERT-base-uncased), and more work is needed to understand if and how our specific results generalize to other masked language models. This is especially important given that past findings comparing gender bias in masked language models with different architectures and model sizes are mixed (e.g., Sharma et al., 2020; Jentzsch and Turan, 2022; Tal et al., 2022).

Additionally, we only considered the task of masked language modeling. We made this choice because psycholinguistic datasets that pair linguistic choices with results of social attitude surveys are rare, and those available to us used language tasks that were most appropriate for evaluation on the task of masked language modeling. However, given that bias on the intrinsic task of masked language modeling may not relate to (extrinsic) bias on downstream tasks (Delobelle et al., 2022), our results (such as BERT's language communicating conservative attitudes) may or may not carry over to downstream tasks. In the future, our approach for relating task predictions to social attitudes could be used to evaluate downstream tasks (such as coreference resolution), once appropriate human data is

available.

Another limitation has to do with differences in the information considered by language models, as opposed to humans, in choosing to use gendered vs. gender neutral language. In both tasks we study, participants and language models evaluate the appropriateness of gender neutral forms based only on contextual cues to the subject's gender, especially gender associations of names. However, when deciding what to say, people can also take into account the referential gender(s) (e.g., the pronouns someone uses, Cao and Daumé III, 2020) of people being referred to. For example, if a person knows that someone named *Michael* uses feminine referential gender, they would likely refer to her with gender neutral or feminine forms (e.g., *congressperson*, *congresswoman*) but probably not masculine forms (e.g., *congressman*). Focusing on evaluation tasks (and language models) which do not consider information about referential gender risks encouraging the development of language technology that performs worse on data from (binary and nonbinary) trans people, and contributing to their erasure. Note that in the *Michael* example there are still linguistic choices (i.e., between *congressperson* and *congresswoman*), which may reflect social attitudes. Future work should study the relationship between linguistic choices and social attitudes in models which can take referential gender into account, while also recognizing the social implications of language choices that respect referential gender.

Finally, while this work developed an approach for evaluating the social attitudes about gender communicated by language models, it does not propose any approaches for improving language models or adjusting the attitudes they communicate. Past work in NLP has discussed different approaches for how pronouns might be handled in language technology (Lauscher et al., 2022), and has developed gender neutral re-writing tasks (Sun et al., 2021; Vanmassenhove et al., 2021), which replace gendered pronouns and words like *fireman/firewoman* with gender neutral variants. Contrasting with standard fairness approaches in NLP that remove information about gender from language technology, work in feminist HCI has discussed approaches for the treatment of gender in language generation which are intended to challenge existing norms and stereotypes, and bring about social change (Strengers et al., 2020). Additionally, work on language reform has discussed the challenges involved in working towards gender-inclusive language, including how explicitly gendered and gender neutral variants can often take on different meanings (Ehrlich and King, 1992; Zimman, 2017). Future work in NLP should consider each of these lines of research, discussing when and how it may be desirable for models to use or represent language that signals gender, and what attitudes those language choices communicate.

7 Ethics and Impact Statement

Because we do not conduct any experiments with human subjects, we are considered exempt from IRB at our institution. The human experimental data we use was previously collected by psycholinguistics researchers for research purposes, and we similarly use it for research purposes only. The experimental stimuli from these datasets do not contain offensive content or information uniquely identifying people, as they consisted of highly controlled (and not offensive) fill-in-the-blank, multiple choice, and sliding scale questions. These datasets also include survey data, which ask about personal information on sensitive topics, such as gender identity. We assume this was taken into account in the IRB process at the institutions where the data was collected.

Licenses for the datasets and models used are in Appendix C. One ethical challenge we encountered in this paper was that, to evaluate BERT on the role nouns dataset, we required frequency counts consistent with BERT's training data, including from English Wikipedia and the BookCorpus (Zhu et al., 2015). However, the BookCorpus is an unlicensed dataset that may violate copyright (Bandy and Vincent, 2021).[12] We were torn between the ethical issue of using an unlicensed corpus, and the scientific issue of needing to use data on which the system we are testing was trained. Based on guidance from reviewers, we decided to present results including (summed) frequencies from both corpora.

The goal of this work is to elucidate connections between language model predictions and social attitudes, focusing on the domain of language and gender. Core to this approach is our focus on psy-

[12] Note that the BookCorpus also contains offensive content and some contact information of authors (Bandy and Vincent, 2021). We did not remove this before collecting frequency counts, as we wanted frequencies that reflect BERT's training data as closely as possible.

cholinguistic data linking language choices and attitude surveys, which sheds light on the ways – both pervasive and nuanced – that language communicates social attitudes. We hope that our approach and results contribute to a broader research agenda examining the attitudes communicated and propagated by language technology, in the context of potential harms and inequities related to gender.

Acknowledgements

We are grateful to Sadie Camilliere, Amanda Izes, Olivia Leventhal, and Daniel J. Grodner for sharing the data used in Section 4. We acknowledge the support of NSERC of Canada (through grants RGPIN-2017-06506 to SS and RGPIN-2019-06917 to BB), and the support of the Data Sciences Institute, University of Toronto (through a Catalyst Grant to SS, BB, and JW).

References

Lauren Ackerman. 2018. Processing singular they with generic and specific antecedents. *Poster presented at AMLaP (Architectures and Mechanisms for Language Processing). Berlin, Germany*, 7.

Kristine M Baber and Corinna Jenkins Tucker. 2006. The social roles questionnaire: A new approach to measuring attitudes toward gender. *Sex Roles*, 54(7):459–467.

Jack Bandy and Nicholas Vincent. 2021. Addressing "documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*.

Nick Barrowman. 2018. Why data is never raw. *The New Atlantis*, (56):129–135.

Connor Baumler and Rachel Rudinger. 2022. Recognition of they/them as singular personal pronouns in coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3426–3432.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Bronwyn M Bjorkman. 2017. Singular they and the syntactic representation of gender in english. *Glossa: a journal of general linguistics*, 2(1).

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How conservative are language models? adapting to the introduction of gender-neutral pronouns. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

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

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Sky CH-Wang and David Jurgens. 2021. Using sociolinguistic variables to reveal changing attitudes towards sexuality and gender. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Kyla Chasalow and Karen Levy. 2021. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 77–89.

Maxwell Roaldseth Davidson. 2014. Development and validation of the transgender prejudice scale.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41(1):87–100.

Susan Ehrlich and Ruth King. 1992. Gender-based language reform and the social construction of meaning. *Discourse & Society*, 3(2):151–166.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Aomputational Ainguistics*, pages 159–166.

Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in BERT–measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*.

Lex Konnelly and Elizabeth Cowper. 2020. Gender diversity and morphosyntax: An account of singular they. *Glossa: a journal of general linguistics*, 5(1).

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. *Proceedings of the 29th International Conference on Computational Linguistics*.

Jack LaViolette and Bernie Hogan. 2019. Using platform signals for distinguishing discourses: The case of men's rights and men's liberation on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 323–334.

Olivia Leventhal and Daniel Grodner. 2018. The processing of gender pronouns and non-binary they: Evidence from event related potentials. Bachelor's thesis, Swarthmore College.

Lucy Li and Julia Mendelsohn. 2019. Using sentiment induction to understand variation in gendered online communities. *Proceedings of the Society for Computation in Linguistics (SCiL)*.

Miriam Meyerhoff. 2014. Variation and gender. *The handbook of language, gender, and sexuality*, 2:87–102.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in nlp: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612.

Brandon Papineau, Rob Podesva, and Judith Degen. 2022. 'sally the congressperson': The role of individual ideology on the processing and production of english gender-neutral role nouns. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in psychology*, page 25.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2020. Evaluating gender bias in natural language inference. In *NeurIPS 2020 Workshop on Dataset Curation and Security*.

Andrew P Smiler and Susan A Gelman. 2008. Determinants of gender essentialism in college students. *Sex Roles*, 58(11):864–874.

Nathaniel J. Smith and Roger Philip Levy. 2008. Optimal processing times in reading: A formal model and empirical investigation. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 595–600.

Elizabeth Stokoe and Frederick Attenborough. 2014. Gender and categorial systematics. *Handbook of language, gender and sexuality*, pages 161–179.

Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the*

4th Workshop on Gender Bias in Natural Language Processing (GeBNLP).

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. Neutral rewriter: A rule-based and neural approach to automatic rewriting into gender-neutral alternatives. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *North American Association for Computational Linguistics (NAACL)*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Lal Zimman. 2017. Transgender language reform: Some challenges and strategies for promoting trans-affirming, gender-inclusive language. *Journal of Language and Discrimination*, 1(1):84–105.

A Appendix: Role Nouns

A.1 Stimuli from Papineau et al. (2022)

Names: The Papineau et al. (2022) stimuli used names selected from among the most popular male and female names (20 each) from the 1998 US Social Security Administration lists, excluding names in the top 100 in both (e.g., Taylor). The male names were Andrew, Austin, Christopher, David, Jacob, John, Joseph, Joshua, Matthew, Michael, Nicholas, and William. The female names were Alyssa, Elizabeth, Emily, Hannah, Jessica, Kalya, Lauren, Madison, Megan, Rachel, Samantha, and Sarah. Note that the Papineau experiment used "Kalya" in place of the name "Kayla" from the Social Security Administration list for female names. To match their experimental stimuli, we used "Kalya" in the sentences we input to BERT.

3-way role noun sets:
anchor, anchorman, anchorwoman
businessman, businessperson, businesswoman
camera operator, cameraman, camerawoman
congressman, congressperson, congresswoman
craftsman, craftsperson, craftswoman
crewman, crewmember, crewwoman
firefighter, fireman, firewoman
flight attendant, steward, stewardess
foreman, foreperson, forewoman
layman, layperson, laywoman
meteorologist, weatherman, weatherwoman
police officer, policeman, policewoman
salesman, salesperson, saleswoman
stunt double, stuntman, stuntwoman

2-way role noun sets:
actor, actress
heir, heiress
hero, heroine
host, hostess
hunter, huntress
villain, villainess

A.2 Determining participant groupings by attitudes

Participant groupings for the role nouns analysis were determined based on responses to the Social Roles Questionnaire from Baber and Tucker (2006). This questionnaire consists of 13 items repeated verbatim here from page 465:

1. People can be both aggressive and nurturing regardless of sex.
2. People should be treated the same regardless of their sex.
3. The freedom that children are given should be determined by their age and maturity level and not by their sex.
4. Tasks around the house should not be assigned by sex.
5. We should stop thinking about whether people are male or female and focus on other characteristics.
6. A father's major responsibility is to provide financially for his children.
7. Men are more sexual than women.
8. Some types of work are just not appropriate for women.
9. Mothers should make most decisions about how children are brought up.
10. Mothers should work only if necessary.
11. Girls should be protected and watched over more than boys.
12. Only some types of work are appropriate for both men and women.

13. For many important jobs, it is better to choose men instead of women.

For each item, participants gave scores indicating numerical values between 0 ("strongly disagree") and 100 ("strongly agree"). For questions 1-5 (the gender transcendence subscale), higher scores indicate more open-minded social attitudes about gender. For questions 6-13 (the gender linking subscale), lower scores indicate more open-minded social attitudes about gender. Using code adapted from Papineau et al. (2022), we subtracted the scores on the gender transcendence subscale from 100, averaged scores per-participant for each subscale, and then took the average of those two values to obtain a final participant attitudes score. Thus, scores range from 0 to 100, with 0 being maximally open-minded about gender. We then selected thresholds so as to create 3 evenly-sized participant groups based on each participant's overall score s: progressive: $s < 12.51$; moderate: $12.51 \leq s < 26.20$; conservative: $26.20 \leq s$.

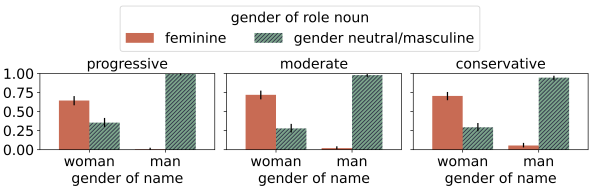
A.3 Frequencies used for $P(V)$ of role nouns

In calculating BERT's posterior probability, $P(V|C)$, of one of the variants of a role noun (Equation (1)), we need frequency estimates for $P(V)$ (the relative frequency of each variant of a role noun set). To as closely as possible match the unigram frequencies that BERT was exposed to, we would like frequencies based on its training corpora, which include Wikipedia and the BookCorpus (Zhu et al., 2015). (The Wikipedia data we use is not exactly the same as the Wikipedia dumps that BERT was trained on, but we assume that the relative frequencies of role nouns are comparable.) The analyses reported in the main body of the paper use the combined frequency counts from these two corpora for each role noun variant.

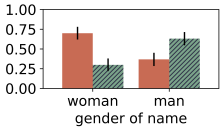
A.4 Correlations of $P(V|C)$ by different methods

To further support the estimated method for calculating $P(V|C)$ using Equation (1) (instead of the direct method of masking and predicting the target role noun variants directly), we carried out the direct method on the 8 role noun sets for which the direct method was possible; 4 each of the 3-way and 2-way role noun sets. (These are role noun sets in which the variants differ in a single word piece.)

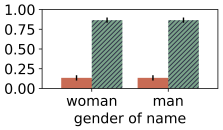
On these 8 role noun sets, we find a strong, significant correlation between the predictions from the direct masking method, and from our method using Equation (1) ($r = 0.76$, $p \ll 0.0001$, $n = 24,000$).[13] Moreover, this is a much stronger correlation than between the direct method and $P(C|V)$ ($r = 0.42$, $p \ll 0.0001$) or $P(V)$ ($r = 0.37$, $p \ll 0.0001$).



(a) Participant responses (Papineau et al., 2022).



(b) BERT predictions



(c) Frequency prior

Figure 3: (a) Participant responses, (b) BERT predictions, and (c) frequency-based predictions, **for 2-way role nouns (e.g., *actress/actor*)** by gender attitudes (progressive/moderate/conservative) and by subject names (women's/men's).

gender attitudes group	loglik for women's names	loglik for men's names	loglik for all data
prog	**-0.82**	-0.95	**-0.89**
mod	-0.88	-0.95	-0.91
cons	-0.88	**-0.92**	-0.90

Table 5: Average log likelihood of data from participants in different gender attitude groups (prog[ressive], mod[erate], cons[ervative]), based on BERT's predictions **(2-way split forms)**. (Num. of observations is 540, 540, and 546, resp.) Higher scores indicate better fit; highest per column (stimulus type) is bolded.

A.5 Details of results on 2-way role noun sets

Participant responses for role nouns with a 2-way split (e.g., *actress/actor*) are shown in Figure 3a, BERT's model predictions on these are shown in Figure 3b, and the frequency baseline predictions in Figure 3c. The average log likelihood per participant group is shown in Table 5.

As with the 3-way role sets, we ask to what extent BERT is sensitive to the same linguistic cues as people – i.e., gender associations with

[13] The n observations are 20 role noun variants (4 3-way and 4 2-way) times 24 names (12 women's and 12 men's) times 50 states.

a woman's or man's name – in making choices among the variants. All participant groups used more FEM variants (e.g., *actress*) for stimuli containing women's names, and more MASC/G-NEUT variants (e.g., *actor*) for stimuli containing men's names. This is captured by BERT's predictions. However, across all participant groups, the non-congruent FEM variants are rarely applied to stimuli containing men's names, and BERT does not capture this result as well. Unlike in the 3-way split analysis, where BERT greatly over-predicted MASC forms for women's names, here BERT over-predicts FEM forms for men's names.

However, looking at the individual items, we find that 2 of the 6 role nouns with a 2-way split – *hostess*/*host* and *heiress*/*heir* – are predicted with high probability to be FEM, for both men's and women's names. Although *hostess* and *heiress* are both less frequent than their MASC/G-NEUT counterparts, our intuition is that both also seem more natural as a bare noun, in sentences like the stimuli here (*NAME is a ____ from STATE*). We hypothesize that the terms *hostess*, referring to a profession, and *heiress*, referring to a social role, can "stand alone", while the MASC/G-NEUT forms are typically used with further specification (of what's hosted, *host at/for [an event]*, or of what's inherited, *heir to [something]*). If BERT's training data reflects these intuitions, then local contextual cues could explain why the FEM forms are predicted with such high probability, even when the stimuli contains men's names.

Our second research question asked which participant group BERT's predictions most resembled. Unlike for the forms with a 3-way split, for the forms with a 2-way split there is much less variation across participants with different gender attitudes, and so differences in BERT's performance are likewise smaller; see Table 5. We find slightly better performance on the progressive group for women's names, which may be due to BERT's high prediction of the (more frequent) MASC/G-NEUT variants, which have been adopted as gender neutral for forms with a 2-way split. However, in general for forms with a 2-way split, BERT's predictions perform comparably (similar log likelihoods) across the different participant groups.

B Appendix: Singular *they*

B.1 Data from Camilliere et al. (2021)

Stimuli used in *they* experiment. The stimulus set included 40 sentence frames with 8 possible critical antecedents (one of each type from Figure 3), plus 15 sentences with singular inanimate controls, for a total of 335 stimuli.

The 40 sentences had a target pronoun evenly distributed across 4 forms of they (*they*, *them*, *their*, *themselves*). We refer to these all as usages of *they*.

The 15 control items each had a singular inanimate noun as the intended antecedent of *they*, as in:

`The cup fell and they broke.`

It was expected that all participants would judge these as unnatural, since singular usage of *they* is valid only for animate antecedents. These items served as controls that allowed Camilliere et al. (2021) to validate that participants who rated *they* as highly natural for all human referents were not simply marking all stimuli as acceptable. As expected, all participant groups gave *they* a relatively low rating when referring to inanimate antecedents (e.g., *the cup*).

Names. Camilliere et al. (2021) assessed gender associations of names based on a norming study from Leventhal and Grodner (2018). The gendered names were Aaron, Adeline, Alice, Amanda, Amelia, Annabella, Bella, Brandon, Bridget, Caleb, Charlotte, Daniel, David, Elena, Elizabeth, Ella, Emily, Emma, Gianna, Grant, Haley, Henry, Isaac, Jacob, John, Joshua, Justin, Lily, Lucas, Maria, Mary, Molly, Nicholas, Penelope, Robert, Scarlett, Vivian, Wyatt, Zach, and Zoey. The non-gendered names were Alex, Cameron, Casey, Dakota, Finley, Frankie, Harper, Hayden, Jayden, Jordan, Justice, Landry, Leighton, Marley, Morgan, Pat, Payton, Remi, Sammy, Skyler, and Taylor.

B.2 Determining participant groupings by attitudes

In addition to the judgments on *they*, Camilliere et al. (2021) had participants complete several surveys, including surveys about acceptance of nonbinary people, familiarity with nonbinary people, Davidson's (2014) Transgender Prejudice Survey, and Smiler and Gelman's (2008) Gender Essentialism Scale. As noted, we drew on the first two – acceptance of nonbinary people and familiarity

with nonbinary people – as Camilliere et al. (2021) found these to be be predictive of higher acceptability ratings for singular *they*. Moreover, we use the nonbinary acceptance score for our grouping of participants on social attitudes because these questions emphasized *attitudes toward* nonbinary people rather than *acquaintance with* nonbinary people. This survey was scored on a scale from 0-5, computed as follows:

- If a person was born female but identifies as male they are a man. +1 if agree
- When I meet someone new I assume that they are either male or female based on what they look like. +1 if disagree
- If someone looks androgynous I try to figure out their gender. +1 if disagree
- People's appearances do not affect what gender pronoun I use to refer to them. +1 if agree
- I think that gender lies on a continuum and is not just male or female. +1 if agree

The range of possible participant values for this scale is relatively small (6 possible values), so we set cut-offs manually, aiming to distribute the scale into roughly evenly sized chunks, while still ensuring enough participants fell into each bin. Participants were grouped based on scores into low acceptance (score of 0), medium acceptance (score of 1-2), and high acceptance (score of 3-5).

B.3 Visualizations with (unadjusted) surprisal

Figure 2 in the main text shows human naturalness ratings alongside predictability according to BERT. In that figure, we used an adjusted surprisal measure to quantify predictability according to BERT, which made it easier to visually compare those results to the human ratings. In Figure 4 here, we present the same plot using unadjusted surprisal, which is a more standard measure. As above, we present results for BERT alongside human ratings.

B.4 Correlations of BERT with participant groups

We carried out additional correlations as in Table 4, between BERT's assessment of *they* and participant ratings, within the different groupings (by linguistic usage stage and by gender attitudes). In all cases, we see the same patterns of a weaker fit to the more progressive participants, in terms of either linguistic stage of usage or gender attitudes.

Grouping of participants ...by linguistic stage	r	p-value
non-innovators	0.60	$p \ll 0.0001$
innovators	0.49	$p \ll 0.0001$
super-innovators	0.26	$p < 0.0001$
...by gender attitudes	r	p-value
low nonbinary acceptance	0.53	$p \ll 0.0001$
med nonbinary acceptance	0.52	$p \ll 0.0001$
high nonbinary acceptance	0.42	$p \ll 0.0001$

Table 6: Correlations between BERT's **raw probability** and mean rating of each participant group, on all 335 stimuli.

Grouping of participants ...by linguistic stage	r	p-value
non-innovators	–0.64	$p \ll 0.0001$
innovators	–0.60	$p \ll 0.0001$
super-innovators	–0.44	$p \ll 0.0001$
...by gender attitudes	r	p-value
low nonbinary acceptance	–0.61	$p \ll 0.0001$
med nonbinary acceptance	–0.63	$p \ll 0.0001$
high nonbinary acceptance	–0.46	$p \ll 0.0001$

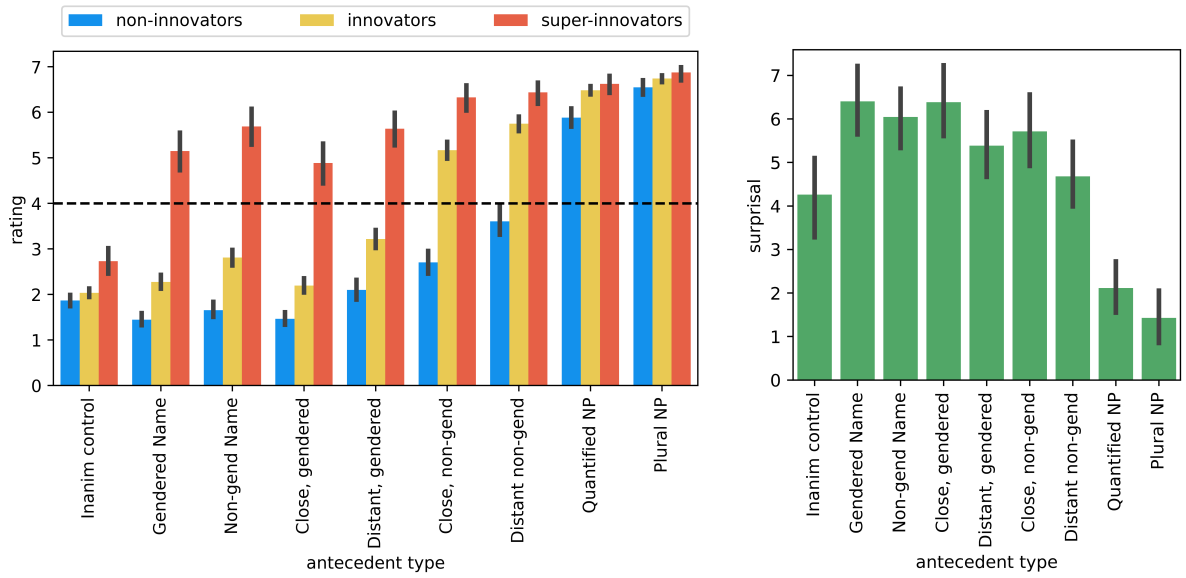
Table 7: Correlations between BERT's surprisal and mean rating of each participant group, on 320 stimuli – i.e., **without inanimate controls**.

In Table 6, we show the correlations for responses of participants grouped by linguistic stage and by gender attitudes, with BERT raw probabilities instead of using surprisal. We aimed to ensure that the pattern of correlations was not changed due to the transform to negative log probabilities instead of directly using BERT's raw probabilities.

In Table 7, we shows the correlations for responses of participants grouped by linguistic stage and by gender attitudes, using only the 8 critical conditions (i.e., removing the inanimate control condition). Here, we wanted to ensure that the pattern of fit to the various groups was not overly influenced by the control condition in which BERT behaved somewhat anomalously compared to people.

B.5 Analysis of attitudes of linguistic participant groups

We aimed to validate that the groupings by linguistic stage, used by Camilliere et al. (2021), reflect social attitudes. To do so, we conduct one-tailed Mann-Whitney U-Tests comparing scores on the nonbinary acceptance survey and nonbinary famil-



(a) Human judgements from Camilliere et al. (2021), by participant group and antecedent type (**higher values mean more natural**)

(b) BERT surprisal (**higher values mean less probable**)

Figure 4: Participant judgements (a) and BERT predictions (b) by antecedent type. Here, BERT predictions are measured using (unadjusted) surprisal (as compared to the adjusted surprisal measure used in the main text). Error bars in all graphs are 95% confidence intervals.

iarity survey across the groups. (Recall, these are the two surveys Camilliere et al. (2021) found to predict ratings on their experimental task.) We find significantly higher (greater acceptance/greater familiarity) scores for the super-innovative cluster than the innovative cluster for both the nonbinary acceptance scale (2.13 for super-innovators vs. 1.27 for innovators, $p = 0.0083$) and the nonbinary familiarity scale (1.25 for super-innovators vs. 0.49 for innovators, $p = 0.0241$). We find no significant differences in survey responses between the innovative and non-innovative clusters on either the nonbinary acceptance scale (1.27 for innovators vs. 1.29 for non-innovators, $p = 0.5805$) or the nonbinary familiarity scale (0.49 for innovators vs. 0.61 for non-innovators, $p = 0.8138$).

C Appendix: Licenses, libraries, and hardware specifications

We use data from Papineau et al. (2022)[14], which is made available under an MIT license. We also use data from Camilliere et al. (2021), which was shared with us directly by the authors.

The model we evaluated was BERT (Devlin et al., 2019), which is released under an Apache License 2.0. The specific model we studied is bert-base-uncased, which has 110 million parameters. We use the PyTorch implementation made available through the HuggingFace Transformers library[15] (library version 4.9.2). All analyses were run on a 2020 M1 MacBook Air; the combined analyses took less than 24 hours of compute time.

We collected unigram frequency counts on the role nouns of Papineau et al. (2022) from Wikipedia and from the BookCorpus (Zhu et al., 2015). For English Wikipedia, we use the dataset made available through HuggingFace[16], which was created based on Wikipedia dumps[17] released under a combination of CC-BY-SA 3.0 and GDFL (unversioned) licenses (data version "20200501.en"). For the BookCorpus dataset, we also use the version available through HuggingFace[18].

We make our code available on GitHub under an MIT license at https://github.com/juliawatson/bert-social-attitudes.

[14] https://github.com/BranPap/gender_ideology

[15] https://huggingface.co/docs/transformers/model_doc/bert

[16] https://huggingface.co/datasets/wikipedia#source-data

[17] https://dumps.wikimedia.org/

[18] https://huggingface.co/datasets/bookcorpus