Remember 'him', forget 'her': Gender bias in the comprehension of pronominal referents

Veronica Boyce (MIT), Titus von der Malsburg (University of Potsdam), Till Poppels (UCSD), Roger Levy (MIT)
vboyce@mit.edu

In rational theories of linguistic communication, utterances convey features of a world model that the speaker has in mind, which the listener reconstructs by integrating utterance content with prior world knowledge and event expectations. In these theories, production and comprehension characteristically are mutually calibrated through Bayesian inference: interpretive preferences reflect the joint influence of prior expectations and speaker preferences for how to linguistically encode a given world-model feature (the likelihood). If, controlling for prior probabilities, form x leads to a stronger inference of world feature a than the inference from form y to world feature b, then a corresponding production preference should hold in: $P(x|a) > P(y|b)$. Here we document a case involving expectations for referent gender and production and interpretation of gendered pronouns in which this calibration between production and comprehension fails to hold: comprehenders treat *he* pronouns as a stronger signal of the referent being male than *she* pronouns are of the referent being female, but there is no corresponding asymmetry in pronoun production preferences.

We report two experiments, both using a set of 80 role nouns with diverse gender stereotypes (ex. *diplomat, butler, nanny, reporter*). In experiment 1, 149 participants each completed 20 passages designed to elicit pronouns referring to a role noun (ex. *The day before the championships, the gymnast worked out. Before and after working out, the gymnast stretched ...*). In experiment 2, 712 participants read researcher-completed versions of the same passages, containing either a *she*, *he*, or *they* pronoun referring to the role noun or a repeat of the role noun (ex. *The day before the championships, the gymnast worked out. Before and after working out, the gymnast stretched her muscles to avoid getting sore.*). After reading 1,4, or 8 vignettes (plus completing a short working memory task), participants were asked to recall the gender of each referent in a forced choice task (options were 'male', 'female', and 'other').

To create an implicit norm, 569 participants completed a gender recall task where all vignettes repeated the role noun instead of using a pronoun. We used this to create an implicit norm by taking the number of 'female' responses divided by the sum of 'male' and 'female' responses. To create an explicit norm, we asked 51 participants to rate role nouns for how male or female they were (ex. *If you encounter a gymnast, how likely are they to be female?*) and created a norm by inverting the 'how male' results, averaging, and rescaling. The implicit norm was highly correlated with the explicit norm (r=.83), but nouns were rated as consistently more male-biased in the implicit norm compared to the explicit norm (Fig. 2).

The pronoun completions from experiment 1 generally track the stereotype of the noun (Fig. 4). However, compared to the explicit norm, we see a bias toward producing *he* more than *she* for neutral nouns. This implies that use of *she* implies more confidence in the femaleness of the referent than *he* implies about the maleness. Using the implicit norm does not reveal a bias and makes *she* and *he* appear equally strong. When it comes to recalling gender, *she* and *he* are not equal. The impact that *he* has on a referent being remembered as 'male' is greater than the impact *she* has on a referent being remembered as 'female' (Figs. 1 and 3). Additionally, we see that as the number of trials increase, the effect of pronoun diminishes, but does not entirely go away (Fig. 1). In this area, *he* is interpreted as being more indicative of maleness than *she* is of femaleness. This is the opposite of what might be expected from the production data.

From this mismatch, we conclude that comprehension and production are miscalibrated with regard to conveying referent gender, in a way that could lead listeners to infer more maleness of referents than the speaker intended to convey. Given how often we use pronouns and role nouns in communication, this distortion is both scientifically interesting and socially troubling.

Sample Stimuli:

Experiment 1: *After the shop on High Street closed for the night, a baker stayed to tidy up. Before the baker took out the trash, [she|he|they|the baker] swept the floor and wiped down the counter.*

Experiment 2: *After the shop on High Street closed for the night, a baker stayed to tidy up. Before the baker took out the trash, . . .*

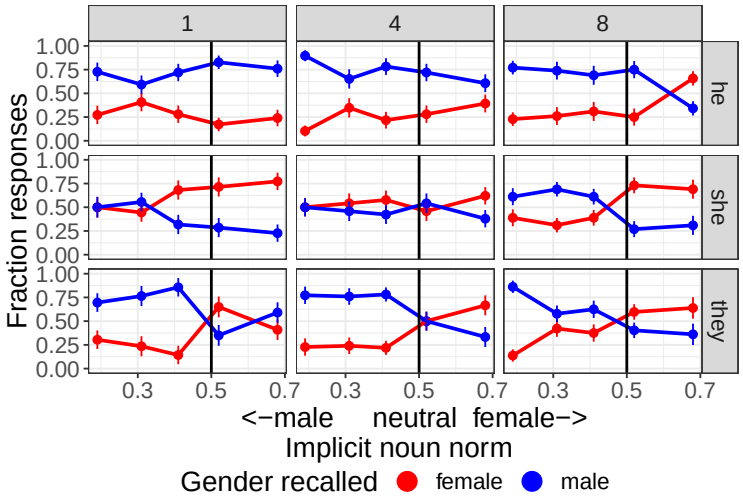

Figure 1: Gender recall responses by trial count, pronoun seen, and implicit noun norm. Comparing *he* and *she*, we see a stronger link between *he* and male responses than between *she* and female responses. For *they*, recalled gender tends to reflect the noun stereotype. Pronoun effects weaken as trial count increases.

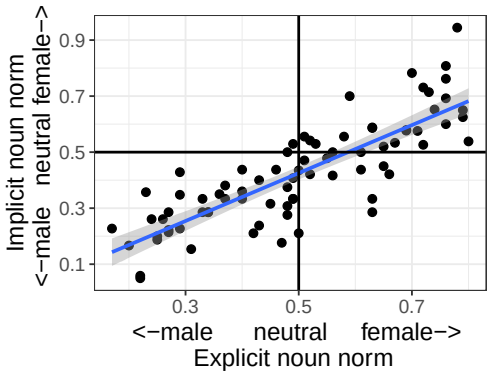

Figure 2: Relation between implicit norm (gender recall in absence of pronoun) and explicit norm (stated gender expectations). The two norms are highly correlated (r=.83), but the implicit norm tends to label nouns as more male.

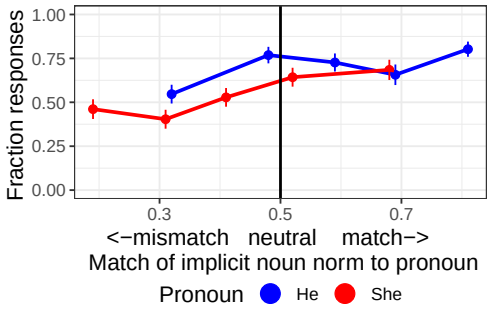

Figure 3: Match of recalled gender to pronoun. The connection between seeing *he* and recalling male is stronger than the connection between seeing *she* and recalling female.

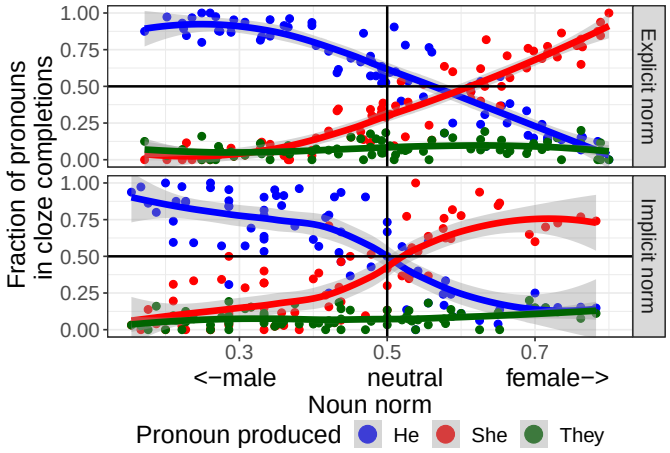

Figure 4: Pronoun completions by noun norms. Using the explicit norm, cloze completions are biased, with more *he* than *she* for neutral nouns, meaning that *she* is a stronger marker of female gender than *he* is of male. Compared to the implicit norm, cloze completions are unbiased, and *he* and *she* are equally strong.