# Gender biases, pronouns and artificial intelligence: Comparing humans and ChatGPT

Elsi Kaiser, Claire Benét Post, Deborah Ho, Haley Hsu and Madeline Rouse

*University of Southern California*

Whether I refer to someone as Ana, Ana Romano or Romano is influenced by many factors. We focus on last-name-only format (e.g. *Schaefer came in,* ex.1a-b). Last-name-only can be used for women (1c), but in many contexts (in U.S.: politics, academia, sports, conversation), it is used more for men (*male bias*, e.g. McConnell-Ginet 2003, Atir & Ferguson 2018). Moreover, researchers referred to by last-name-only are judged more famous and deserving of awards (*eminence bias*, Atir & Ferguson 2018).

We have two aims. First, we ask whether the male and eminence biases emerge even in controlled contexts where speakers lack rich mental representations of the referents and if the male bias can override strong semantic verb biases. Second, we test if ChatGPT, a chatbot based on large language models (LLMs), has the same biases as humans. The *explicit* biases of LLMs have received a lot of attention (e.g. Borji 2023, Doshi et al. 2023, Walther et al. 2023) and are being addressed (e.g. Borji 2023), but it is unknown whether LMMs like ChatGPT exhibit *implicit effects* (e.g. linking last-name to male/eminence) which may be harder to address within the model architecture. This has practical implications (e.g. avoiding bias) and theoretical ramifications (e.g. how human develop/acquire biases).

In **Exp.1** (20 targets, 22 fillers, 91 native US-English speakers), people read items ending in '*because+pronoun'* and wrote continuations. We manipulated: **(i)** verbs' implicit causality (IC): When followed by an explanation signaled by *because*, does the verb elicit subject (IC1) or object (IC2) continuations? (Verbs selected based on Hartshorne & Snedeker 2013; see also Bott & Solstad 2021); **(ii)** whether the pronoun is *he* or *she*; **(iii)** whether the verb is eminent (presents IC-biased referent in a positive light, e.g. IC1: *impressed*, IC2: *promoted*) or noneminent (negative light, e.g. IC1: *disappointed*, IC2: *despised*). Targets (ex.2-3) had one first and one last name, with the last-name in the IC-favored position. Thus, *she* conditions pit verb bias against the male bias of last-name-only format.

**Results Exp.1**: Fig.1 shows how often people use pronouns for the last-name-only referent favored by IC bias (subj/IC1 verbs, obj /IC2 verbs). All *he* conditions show IC effects in the expected directions (p's<.001), replicating prior work on implicit causality. But none of the *she* conditions show IC effects. With *she*, even when IC favors the last-name-only referent (*Smith impressed Amanda because she*), people are reluctant to interpret the last-name referent as the antecedent of *she*, despite verb bias: There is a ***male bias with last-name format*** (with both male and female participants). What about the ***eminence bias***? With IC2 verbs, IC effects are weaker with non-eminent objects with both *he* and *she* (p's<.02). We attribute this to an eminence-related reluctance to provide explanations of why a last-name-referent would be criticized, despised etc. IC1 verbs show no such effects, maybe due subject topicality/salience.

In **Exp.2**, ChatGPT continued all 20 targets 40 times (40 'participants'). Fig.2 shows ChatGPT has a stronger eminence bias and a weaker male bias. With eminent verbs (left 4 bars), both *he* and *she* elicit mostly verb-bias compatible interpretations (p's<.01), unlike Exp.1 where *she* conditions did not differ from chance: **ChatGPT's male bias is weaker** and *it easily interprets even she as referring to a last-name referent when that referent is favored by verb bias*, though still at rates below *he*. With non-eminent verbs, ChatGPT yields fewer verb-bias-compatible continuations with both *he* and *she* (p's<.01; rightmost 3 bars <30%): a **strong eminence bias**. Like Exp.1, this is clearer with IC2 than IC verbs.

In sum, last-name-only format with humans shows a **male bias**, strong enough to counteract well-known IC verb biases. We also find an eminence bias. In contrast, ChatGPT shows a **stronger eminence bias and a weaker male bias**. This may seem unexpected, given criticism of ChatGPT's gender biases. It may be that the male bias is not as easily learnable from linguistic input (ChatGPT training data) as the eminence bias, and may only emerge in conjunction with social, extra-linguistic experience.

(1a) “I would go so far as to say that had *Watson* and *Crick* not come into Rosalind's photograph -- by hook or crook; whichever way it was -- they would have lost the race entirely” (from podcast by Scientific American on Dr. Rosalind Franklin)
(1b) “*Johnson* is a great professor. He is funny” (from ratemyprofessor.com)
(1c) “*Welsh* is my favorite professor. She's just amazing” (from reddit.com)

(2) **Example item: IC1 verb (subject-biased)**

| | |
|---|---|
| (a) Smith impressed Eric because he… | [he + eminent verb] |
| (b) Smith impressed Amanda because she… | [she + eminent verb] |
| (c) Smith disappointed Eric because he… | [he + non-eminent verb] |
| (d) Smith disappointed Amanda because she… | [she + non-eminent verb] |

(3) **Example item: IC2 verb (object-biased)**
Frank {promoted/despised} Mayfield because {he/she}…[he/she + eminent /non-eminent verb]

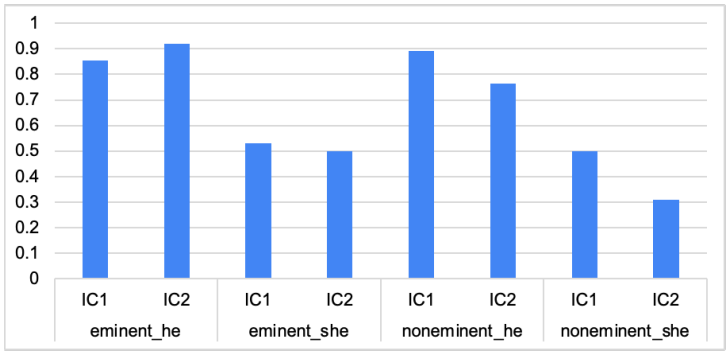


**Fig.1** Exp.1: Y-axis: Proportion of continuations compatible with verb’s IC bias (IC1 verbs: pronoun refers to subject, IC2 verbs: it refers to object). Data were double-coded by coders blind to condition.

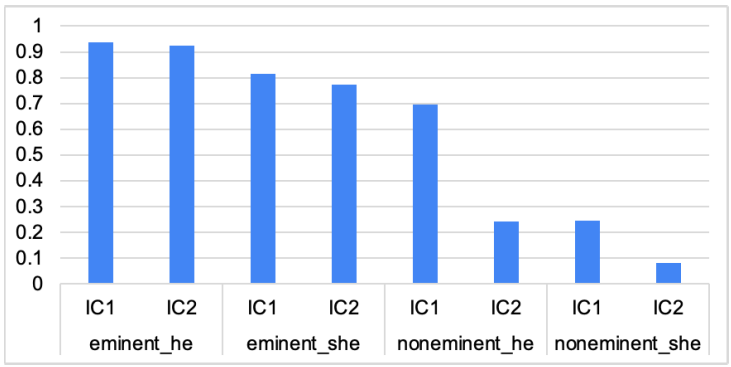


**Fig.2** Exp.2 (ChatGPT, January 9, 2023 version): Y-axis is the same as in Fig.1

**References**

Atir, S., & Ferguson, M. J. (2018). How gender determines the way we speak about professionals. *Proceedings of the National Academy of Sciences,* 115(28), 7278-7283.
Borji, A. (2023). A categorical archive of ChatGPT failures. arXiv preprint arXiv:2302.03494.
Bott, O., & Solstad, T. (2021). Discourse expectations. *Linguistics*, 59(2), 361-41
Doshi, R., Bajaj, S. & Krumholz, H. (2023). ChatGPT: Temptations of Progress. *A J of Bioethics*, 1-3.
Hartshorne, J. & Snedeker, J. (2013). Verb argument structure predicts implicit causality. *Language and Cognitive Processes,* 28, 1474-1508.
McConnell-Ginet, S. (2003). “What's in a name?” Social labeling and gender practices. *The handbook of language and gender*, 69-97.
Walther, A., Logoz, F., & Eggenberger, L. (2023). *The Gendered Nature of AI: Men and Masculinities Through the Lens of ChatGPT.*