Referring to people with only a last name: Comparing gender biases in humans and chatGPT

Elsi Kaiser
University of Southern California, Los Angeles, California, United States

Deborah Ho
University of Southern California, Los Angeles, California, United States

Haley Hsu
University of Southern California, Los Angeles, California, United States

Claire Benet Post
University of Southern California, Los Angeles, California, United States

Madeline Rouse
University of Southern California, Los Angeles, California, United States

Abstract

In some contexts, last-name-only format is used to refer to people (e.g. "Hedberg came in", "Jones called"). At least in U.S. English, men are more likely to be referred to by last-name-only format than women (male bias, e.g. McConnell-Ginet 2003). Moreover, researchers referred to with last-name-only are judged more famous/eminent (eminence bias, Atir & Ferguson 2018). However, the robustness of these biases is not yet well-understood, nor how they interact with other semantic biases. We report sentence-completion data from humans showing that these biases persist in informationally-impoverished contexts, the male bias persists when even pitted against verbs' implicit-causality biases, and these biases persist even when use of the last-name-only format for women is primed. Furthermore, we compare the human results to sentence-completions produced by the language model chatGPT, and show that chatGPT exhibits a weaker gender-bias effect but a stronger eminence bias than human participants.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*