“Did You Really Mean That?”: Stereotyped Inferences Regarding Gender and Social Characteristics from Dialogue Fragments between a Human and Conversational Assistant (CA)

Amelia Cavaness[1], Libby Ferland[2], Wilma Koutstaal[1]

[1]Department of Psychology, [2]Department of Computer Science and Engineering

Definitions

Non-stereotyped inference: an inference based on individualized or contextual information without generalizing based on group traits.

Stereotype inference: a generalized belief about behaviors, traits or roles based on a group identity (e.g., gender, occupation, ethnicity) which can be either explicit or implicit.

Research Questions

1. Do stereotypes more often pertain to certain mental domains (i.e., emotional state, personality, habits, biographical/factual, morals and values, other)?
2. Are humans more likely than GPT to generate stereotypical inferences, and how do these differ between humans and GPT?

Introduction

▪ **Human-Human Interactions and Stereotyping**

- Stereotyping: unconscious and automatic cognitive shortcut that can simplify social interactions, but lead to biased assumptions, contributing to prejudice and discrimination.
- Social Role Theory: Gender differences in traits and behaviors arise from societal roles, not inherent biological differences[4].
- Gender Schema Theory: Children learn gender-appropriate behaviors through socialization, forming mental schemas that guide their perceptions and behaviors[1].
- Stereotype Activation: Social role names (e.g., “nurse” or “surgeon”) can trigger gendered inferences, even in ambiguous contexts[7].

▪ **Human-CA Interactions and Stereotyping**

- Training Biases: CAs trained on biased data may perpetuate harmful stereotypes.
- User Influence: Users may influence CAs to reinforce stereotypes through their own biased language, creating a feedback loop[2].
- Gendered Personalities: Virtual assistants (e.g., Siri, Alexa) may reinforce societal gender norms, portraying women as assistants[5].
- Cultural Biases: CAs may fail to recognize or misrepresent non-western cultures, especially if trained on predominantly English-language data[6].

▪ **Study Overview**

- Exploratory Study: Assess the frequency of stereotype-driven inferences from human participants and GPT on brief dialogue snippets.
- GPT Variants: GPT was presented dialogue snippets with or without persona prompts (age 22, gender M/F).
- Research Focus: Examine how subtle cues activate gender and social characteristic schemas, reinforcing stereotypes.

Selected References

[1] Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological Review, 88*(4), 354–364. https://doi.org/10.1037/0033-295X.88.4.354

[2] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

[3] Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies. *Proceedings of the 1st International Conference on Intelligent User Interfaces - IUI '93*. https://doi.org/10.1145/169891.169968

[4] Eagly, A. H. (1987). *Sex differences in social behavior : a social-role interpretation.* Lawrence Erlbaum Associates.

[5] Habler, F., Schwind, V., & Henze, N. (2019). Effects of smart virtual assistants' gender and language. *Proceedings of Mensch Und Computer 2019 on - MuC'19*. https://doi.org/10.1145/3340764.3344441

[6] Peters, U., & Carman, M. (2024). Cultural bias in explainable AI research: A systematic analysis. *Journal of Artificial Intelligence Research, 79*, 971–1000. https://doi.org/10.1613/jair.1.14888

[7] Reynolds, D., Garnham, A., & Oakhill, J. (2006). Evidence of immediate activation of gender information from a social role name. *Quarterly Journal of Experimental Psychology, 59*(5), 886–903. https://doi.org/10.1080/02724980543000088

Method

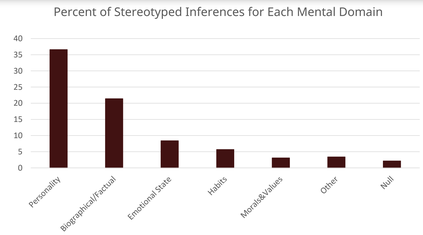
Phase 1: N = 33 participants attended four CA-interaction sessions during a two-day period. Participants responded to a CA prototype that had been controlled by a researcher in a separate room via a “Wizard-of-Oz” set-up[3]. The CA asked participants questions about their schedule, habits and corresponding stress levels for their daily activities. “Stimulus pairs” made up of a prompt or question from the CA and a response from the participant were extracted.

Phase 2: A different group of participants (N = 37) were then presented these dialogue fragments in written form and asked to generate three inferences that they could make about the human speaker based on the provided text.

Phase 3: The large number of generated inferences from Phase 2 were rated by two research assistants on a set of further dimensions, including whether the instance involved self-disclosure, its context, its mental domain (or “form”), valence, level of intimacy, relatedness, and inferences about the speaker. One of the two research assistants then further independently assessed whether or not the inferences contained a stereotype based on the correlating dialogue fragment. The research assistants were initially unaware that some of the inferences were generated by GPT that had been given different gender and age-related persona to generate the inferences.

Results

- **Overall Frequency of Stereotypes**
  - Total Inferences: 4,552 generated by participants
  - Stereotyped Inferences: 383 (8.4% of total)
- **Frequency of Stereotypes Generated by Human vs. GPT**
  - GPT-Generated: 28 (2% of GPT’s 1,364 total inferences)
  - Human-Generated: 355 (11% of humans’ 3,198 total inferences)
- **Impact of GPT Personas on Stereotype Generation**
  - GPT Personas with Gender/Age: 18 out of 28 stereotyped inferences (female, female+age22, male, male+age22)
  - Default GPT (no specified persona): 10 stereotyped inferences



Mental Domain	Speaker 1	Speaker 2	Human Inference
Personality	“After the project meeting you planned to call a friend at around five. How did it go?”	“It went really well. I called my friend; her name is [FRIEND’S_NAME]... and we talked for thirty minutes.”	She’s extroverted.
Biographical/Factual	“How long will you sleep?”	“Hopefully at least twelve hours <laugh>, I don’t have to get up tomorrow morning.”	They have sleep problems.
Emotional State	“How stressful do you think going to Sonic will be?”	“A two because sometimes the line is long and I don’t like that.”	The speaker gets anxious in social situations.
Habits	“Do you usually cook?”	“...I do once a week, normally, I help my mom.”	When she does (cook), it’s with her mom.
Morals and Values	“Good afternoon {NAME}. How are you this afternoon?”	“Umm right now I’m sweaty ‘cause I just worked out. How are you?”	Fitness is one of the person’s few interests.
Other	“What was happening in [NEXT_CLASS_NAME]?”	“Ah nothing, there was just some concepts that the professor went over that... were really unclear at times.”	The professor was boring.
Null	“How did it [group project] go?”	“One person didn’t show up... so bad.”	They prefer to work alone.

Examples of the stimulus pairs and instances of stereotyping by humans, separately for the several different mental domains in Form 1.

Discussion & Conclusion

- **Mental Domains of Stereotype Inferences**
  - Overall Stereotyping: 8.4% of inferences were rated as containing a stereotype, indicating that stereotypical thinking was present but not dominant.
  - Unclear Expectations: Uncertainty whether this frequency is higher or lower than expected, considering limited context and task instructions.
  - Majority of Inferences: Participants did not heavily rely on gender or social stereotypes, focusing on other types of information in most cases.
- **Primary Mental Domain (Form 1)**
  - Most Common Domain: Personality (36.6%), suggesting societal influences like Social Role Theory[4], on gendered communication traits.
  - Second Most Common Domain: Biographical/ Factual (21.4%), similar to both Gender Schema Theory[1] and Social Role Theory[7] in that inferences are made based on characteristics like gender, age, or occupation.
  - Less Common Domains: Emotional State, Habits, Morals/Values.
- **Null Responses**
  - 21% in Form 1 and 89% in Form 2; many inferences lacked clear categorization in Form 2, suggesting complexity in participants’ thinking.
  - Ambiguity and Complexity: Some inferences were difficult to categorize, possibly due to mental fatigue or unclear reasoning by raters.
  - Rater Challenges: Some inferences were not easily explained, leading to “null” categorizations.
  - Higher proportion of Form 2 null responses may indicate that a primary domain was already sufficient.
- **Human vs. GPT Stereotyping**
  - Human-Generated Inferences: human inferences were stereotyped substantially more often than GPT-generated inferences.
  - Human Cognition: Humans may rely on stereotypes as cognitive shortcuts, especially with limited information[7].
  - Implication: GPT’s design may mitigate stereotypical thinking, especially when not prompted with a persona.
- **Limitations & Future Directions**
  - Sample Limitations: The participant sample in Phase 2 lacked diversity (mostly female university students). Findings may differ with more diverse demographics or languages.
  - Rater Reliability: Two research assistants rated inferences; a larger group or consensus-based approach could reduce bias and improve reliability in interpreting difficult inferences.
  - Future Studies: Expand the participant pool, include more raters and consider cultural diversity to enhance findings.