Misplaced Capabilities: Evaluating the Risks of Anthropomorphism in Human-AI Interactions

Takuya Maeda

Western University
tmaeda@uwo.ca

Abstract

The present research examines anthropomorphism, or human-like features, in conversational AI systems as a design element that facilitates human-AI interactions. The paper outlines how these human-like features intersect with user perceptions in ways that can co-create misplaced trust, and it explores ways to de-anthropomorphize AI systems. Using role-based prompts to elicit different anthropomorphic features within chatbot language and design, the study identifies and categorizes different types of anthropomorphism exhibited by large language models (LLMs), a necessary step towards evaluating the appropriate use of such features in technical systems. The role-based prompting process also provides a way to explore the stability of LLM responses. Ultimately, the paper explores how this approach could be incorporated into user studies to understand users' motivations, and it discusses the need for design interventions that can mitigate harms and biases hidden in human-AI interactions.

Introduction

As large language models (LLMs) become increasingly sophisticated and human-like in their communication across various modalities, it is essential to examine the social and design implications of their use in information retrieval. This paper proposes a novel framework for evaluating the potential risks posed by anthropomorphic features in LLMs, focusing on harms that arise not only from algorithmic biases in machine learning models or data, but also from human-computer interactions themselves (Gabriel et al. 2024; Weidinger et al. 2022). In the context of chatbots, user behavior plays a critical role in shaping how predictive outputs are interpreted by recreating social contexts around these responses. This dynamic can introduce stereotypes and biases that affect how users appraise and utilize generated information. For example, research shows that people often apply gender and racial stereotypes to computers without critical reflection (Nass and Moon 2000; Abercrombie et al. 2021), reinforcing harmful representations and influencing perceptions of trust and persuasiveness (Ruane, Birhane, and Ventresque 2019). Understanding human-AI interactions as parasocial relationships—one-sided interactions between users and conversational agents—provides a lens to investigate how chatbot design and user behavior contribute to algorithmic harms. Anthropomorphism in these systems is both "enacted" by chatbots through design choices and projected by users based on their perceptions and expectations. This perspective also highlights an essential aspect of chatbot design: the deliberate creation of artificial intimacy to foster user engagement (Abercrombie et al. 2023).

Research Goals and Questions

In order to understand how to best evaluate anthropomorphism in chatbots, we must first acknowledge how anthropomorphism is collaboratively formulated by users and chatbots, respectively. This involves answering two questions: (1) What are the different types of anthropomorphic features embedded within LLM responses and chatbot design, and what kinds of interactions do they stimulate in users?; (2) What are the different social roles that users assign to chatbots, and how consistently do chatbots embody them? More specifically, how do these assigned roles affect generated responses? Do role-specific responses exhibit more anthropomorphic language? Do they exhibit more socio-cultural biases?

Previous Work

My previous work has focused on defining anthropomorphism within the context of human-AI interactions, developing a conceptual framework that integrates theories of parasociality, social affordance, and trust. This framework explores the mediated, interactive spaces between users and AI systems and the modes of interaction that govern them (Maeda and Quan-Haase 2024). It identifies key mechanisms, such as affirming language and black-box design, which collectively create an illusion of reciprocal engagement between users and chatbots. These interactions often involve what I, following Stark (2024), term "projective inference": a one-sided dynamic where users impose their own agency, experiences, and biases to contextualize chatbots' predictive outputs. Such user behavior can transform predictive outputs into seemingly social responses, such as answers or advice, effectively leading users to assign social roles to AI systems. This framework also highlights the potential harms of parasocial interactions with anthropomorphized chatbots, including role displacement, misaligned

tasks, cognitive priming, and emotional attachment. These risks often encourage users to rely on pathos rather than reason, overlooking critical issues in AI-generated information. I attribute these phenomena to the concept of "parasocial trust," a unique, one-sided form of trust that emerges in human-AI interactions. This trust amplifies the perceived reliability and intimacy of chatbots, even in contexts where such perceptions may lead to ethical and practical concerns.

Approaches

Based on my conceptual framework, I examine how LLM responses are shaped by specific social roles assigned through role-based prompts. An example of a role-assigned prompt is: "You are [ROLE] based in [Country, City, etc.]. Please generate [scripts, characters, profiles, ideas, etc.]." Building on the work of Inie et al. (2024), which categorizes linguistic features that signal cognition, biological metaphors, and agency, my research aims to develop a taxonomy of anthropomorphic design features. This taxonomy focuses on relational cues such as sympathy ("It's completely okay to feel tired and unmotivated"), reciprocity ("So what do you think?"), and encouragement ("You got this!"). Role-based prompts allow for the categorization of various tones and content in LLM outputs, uncovering different forms of human-like expressions. By comparing outputs generated across different roles and personas, this approach highlights how responses might reflect underlying assumptions present in the model's training data and interpretative mechanisms, which may be tied to specific contexts. Variables include professional roles (e.g., tutors, engineers), personal roles (e.g., friends), task types (e.g., editing, coaching, brainstorming), emotional cues (e.g., struggle), and sociocultural contexts (e.g., day-to-day interactions, habits).

Future Directions

The motivation for conducting lexical analysis is not only to elucidate biases embedded in assigned or embodied social roles but also to inform future user studies. Results related to relational and emotional responses could help design tasks and interactive scenarios that simulate various types of human-AI relationship dynamics. Analyzing chatbot functions through the lens of assigned social roles offers a novel means to evaluate chatbot capabilities. It can also help illuminate users' motivations—specifically, how and why people rely on chatbots for specialized tasks, such as mental health support, rather than consulting professionals. It may also shed light on why users form unhealthy relational bonds with chatbots, exploring factors such as emotional vulnerability, unmet social needs, or the anthropomorphic design features that foster artificial intimacy. Aligning users' expectations with desirable traits in LLM responses is challenging, underscoring the need to explore psychological factors that facilitate meaningful and reciprocal interactions. The insights derived from these explorations could help identify which aspects of anthropomorphism are acceptable in different contexts and what strategies should be employed to manage human-AI interactions effectively. For instance, when should harms be mitigated by de-anthropomorphizing chatbots, and when should user literacy about AI capabilities and limitations be prioritized? Addressing these questions would enable the establishment of realistic expectations for chatbot performance and the development of practical guidelines for their use. Ultimately, this research aims to (1) enhance our understanding of human-AI interactions, (2) introduce a novel framework for evaluating the social and design implications of conversational AI systems, and (3) clarify the appropriate application of human-like features to facilitate safe and effective human-AI interactions. By doing so, this work can contribute to more responsible AI development.

References

Abercrombie, G.; Cercas Curry, A.; Dinkar, T.; Rieser, V.; and Talat, Z. 2023. Mirages. On Anthropomorphism in Dialogue Systems. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4776–4790. Singapore: Association for Computational Linguistics.

Abercrombie, G.; Cercas Curry, A.; Pandya, M.; and Rieser, V. 2021. Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. In Costa-jussa, M.; Gonen, H.; Hardmeier, C.; and Webster, K., eds., *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 24–33. Online: Association for Computational Linguistics.

Gabriel, I.; Manzini, A.; Keeling, G.; Hendricks, L. A.; Rieser, V.; Iqbal, H.; Tomašev, N.; Ktena, I.; Kenton, Z.; Rodriguez, M.; et al. 2024. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244*.

Inie, N.; Druga, S.; Zukerman, P.; and Bender, E. M. 2024. From" AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2322–2347.

Maeda, T.; and Quan-Haase, A. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1068–1077. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Nass, C.; and Moon, Y. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1): 81–103.

Ruane, E.; Birhane, A.; and Ventresque, A. 2019. Conversational AI: Social and Ethical Considerations. *AICS*, 2563: 104–115.

Stark, L. 2024. Animation and Artificial Intelligence. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1663–1671. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.