Let's Chat: Understanding User Expectations in Socialbot Interactions

Elizabeth Soper* and **Erin Pacquetet*** and **Sougata Saha†** and **Souvik Das†** and **Rohini Srihari†**

SUNY at Buffalo, Departments of Linguistics* and Computer Science†
{esoper,erinmorr,sougatas,souvikda,rohini}@buffalo.edu

Abstract

This paper analyzes data from the 2021 Amazon Alexa Prize Socialbot Grand Challenge 4, in order to better understand the differences between human-computer interactions (HCI) in a socialbot setting and conventional human-to-human interactions. We find that because socialbots are a new genre of HCI, we are still negotiating norms to guide interactions in this setting. We present several notable patterns in user behavior toward socialbots, which have important implications for guiding future work in the development of conversational agents.

1 Introduction

In recent years, it has become increasingly common for humans to interact with computers through natural language, either through speech (e.g. voice assistants) or through text (e.g. customer service chatbots). Most of these interactions have a specific functional goal; users may ask a bot to perform tasks such as giving the weather forecast, setting a timer, or making a dinner reservation. It is less common for users to engage in purely social conversations with a bot – chit-chat remains a primarily human mode of language.

In this paper, we explore data collected during the Alexa Prize Socialbot Grand Challenge 4[1] (Ram et al., 2018; Khatri et al., 2018), where teams designed chatbots to have social 'chit-chat' conversations with humans, with the goal of mimicking human interactions. Users conversed orally with socialbots via an Alexa-enabled device. We analyze this data in order to better understand user behavior: how do the human-bot interactions differ in nature from typical human conversation? What are users' expectations of a socialbot, and how can we develop socialbots which better meet these expectations? The human-centered analysis of socialbot interactions presented here aims to inform future research in developing natural and engaging conversational agents.

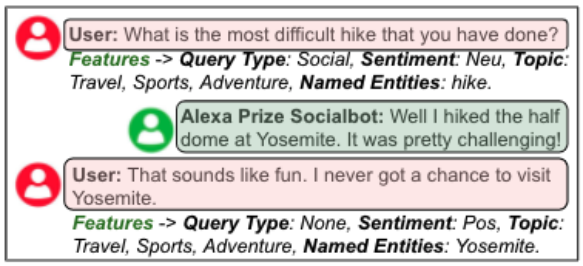

Figure 1: Sample conversation with annotated features.

Of course, the quality of the bot's responses plays an important part in how the user interacts with it; if the bot's responses aren't human-like, users won't treat it like a human. In this paper, our primary goal is not to evaluate the quality of this particular socialbot, but rather to get a sense of what users want from socialbots in general. Once we understand user expectations, we can design socialbots which better satisfy these expectations. The rest of this paper is organized as follows: in §2 we summarize previous work studying conversation, in both human-to-human and HCI settings. Next, we analyze new Alexa Prize data: in §3 we describe ways in which users treated the bot the same as a human, and in §4 we highlight ways that users behave differently with the bot than they would with a human. We discuss the implications of our analysis in §5, and finally conclude in §6.

2 Previous Work

There is a long tradition of literature studying the social and linguistic rules of human discourse. H.P. Grice, in particular, formalized many of the underlying assumptions that we make when conversing with humans. His *cooperative principle* holds that speakers must work together to negotiate the terms of a conversation (Grice, 1989). He further breaks this principle down into four maxims of conversation (quantity, quality, rela-

*Equal Contribution

[1]https://www.amazon.science/alexa-prize/socialbot-grand-challenge/2020

Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 34 - 39

tion, and manner) which specify the assumptions required for cooperative conversations. Other work has also highlighted the importance of established scripts for different scenarios (Hymes, 1972; Tomkins, 1987).

The history of research on HCI is shorter but vibrant. Early work questioned how we should conceptualize AI, and made predictions about how more human-like computers might fit into our lives (Mori, 1970; Winograd et al., 1986). As conversational agents became more widespread, these predictions have been put to the test, with two major patterns surfacing:

The first pattern is that humans tend to treat computers as if they were humans. The Computers as Social Actors (CASA) paradigm (Nass and Moon, 2000) holds that people will "mindlessly" apply existing social scripts to interactions with computers. In an early study of HCI, Nass and Moon (2000) showed that people demonstrated politeness and applied gender stereotypes to computers, even though they were aware that such behavior didn't make sense in the context. Posard and Rinderknecht (2015) show that participants in a trust game behaved the same toward their partner no matter whether they believed the partner to be human or computer. Such results support the idea that humans tend to apply existing social scripts to computers, even when they are aware that the scripts may not make sense for the situation.

Assuming that user expectations of computers are identical to their expectations of humans may be overly simplistic, however; in other studies of HCI, a different pattern emerges. Mori (1970) posited that increased humanness will increase a computer's likeability up to a certain point, past which it will become 'uncanny' or creepy, a phenomenon which he dubs *The Uncanny Valley*. The Uncanny Valley of Mind theory holds that people are uncomfortable with computers that seem too human. Gray and Wegner (2012) find that computers perceived to have experience (being able to taste food or feel sad) are unsettling, whereas computers perceived to have agency (being able to retrieve a weather report or make a dinner reservation) are not. Clark et al. (2019) found in a series of interviews that users have different priorities in conversing with computers versus other humans. Shi et al. (2020) found that people were less likely to be persuaded to donate to a charity when they perceived their interlocutor to be a computer. Other recent studies have found similar differences in interactions with virtual assistants (Völkel et al., 2021; Porcheron et al., 2018). All of this evidence suggests that, while people may default to existing social scripts in interacting with computers, they may not be comfortable treating a computer identically to a human.

In this paper, we extend the existing literature on HCI to a new genre by analyzing user interactions with socialbots. We find evidence that users "mindlessly" apply social rules and scripts in many cases (see §3) as well as evidence that users adapt their behavior when conversing with the social bot (see §4). Overall, we conclude that although socialbots are designed to mimic human interactions, users have fundamentally different goals in socialbot conversations than in typical human conversation, but that the norms of socialbot interactions are still being actively negotiated.

3 Dataset

We analyze a subset of the live conversations collected by one of the finalists of Alexa Prize 2021 (Konrád et al., 2021; Chi et al., 2021; Saha et al., 2021; Walker et al., 2021; Finch et al., 2021). The dataset comprises 8,650 unique and unconstrained conversations conducted between June and October 2021 with English-speaking users in the US. With a total of 346,554 turns and an average of 44 turns per conversation, the dataset is almost twice the size of the existing human chat corpus ConvAI (Logacheva et al., 2018). Further, with a ratio of 1.1 conversations per user, the corpus significantly exceeds the number of unique users, compared to similar previous studies (Völkel et al., 2021; Porcheron et al., 2018; Völkel et al., 2020). The dataset also contains user ratings measuring conversation quality on a Likert scale from 1 to 5, making it possible to analyze the impact of diverse conversational features on overall user experience. Fig. 1 depicts a sample conversation, along with some of the features.

4 How do users treat the socialbot like a human?

Conversation can serve two broad purposes: social and functional. Social conversations aim to build a rapport between the interlocutors, whereas functional conversations aim to achieve some practical goal. Clark et al. (2019) found that this dichotomy was important in explaining differences between

human-to-human and human-computer conversation; their participants found social conversation less relevant when interacting with computers.

We manually identified salient phrases for each conversation type (social vs. functional) from a subset of the conversations, and found that 65% of user queries are social in nature; these queries include seeking opinions, preferences, and personal anecdotes (see Appendix Fig. 2). This shows that, contrary to the findings of Clark et al. (2019), socialbot users actually engage in social conversation more than purely functional conversation. This suggests that the preference for functional conversation reported in Clark et al. (2019) is situational in nature, rather than a general preference in human-computer interactions.

Another way that user behavior towards the socialbot mimics human conversation is the use of indirectness. Around 21% of the socialbot's Yes/No queries result in a user response which does not include *yes* or *no*. In these cases, the bot must infer the connection between the question and the user's answer as in (1), where the user's answer implies *no*.

(1) BOT: "Whenever I have a craving, I order food online from my favorite restaurant. Do you?"
USER: "I do drive through."

Making the necessary inferences to understand and appropriately respond to such indirect responses is quite difficult for conversational agents, but users assume that the bot can follow their implicatures as easily as a human would. This evidence seems to support the CASA theory, showing that humans mindlessly apply human expectations to the bot.

5 How do users treat a bot differently from a human?

While a surprisingly high proportion of user queries are social in nature, that leaves 35% of queries that are functional in nature, including requests for the bot to perform a task (*Can you sing please?*), or provide information (*Who directed Jurassic Park?*). While not as frequent as social queries in our data, functional queries are still much more common than would be expected in human conversation. Functional queries generally lead to higher ratings on average than social queries (see Appendix Fig. 2 for a detailed breakdown). This suggests users' preference for functional interactions with computers. This could also be explained by the bot performing better in a functional mode than social, or by preconceived user expectations from interactions with other bots. However, although this socialbot will answer factual questions, it does not act as a smart assistant and will reject requests to perform Alexa-assistant commands.

Another clear difference between socialbot and human conversations is the violation of traditional Gricean maxims. As is customary in the US, the socialbot begins by asking the user how they are doing. In human conversation, this question is almost invariably followed by some form of "I'm fine. And you?" Such phatic conversational openings serve to establish a rapport between speakers. By contrast, in the socialbot data we find that in 9.3% of cases, the user disregards this greeting and starts a new topic, as in (2).

(2) BOT: "Hi. How's your day going so far?"
USER: "Do you want me to tell a joke?"

We find this type of abrupt shift also happens beyond the initial "How are you?" exchange. Users don't feel obligated to obey the Gricean maxim of relevance by responding directly to queries, as they would in human conversation, because the bot is programmed to respond to any queries and try to continue the conversation. Using high-precision keyword-based mappings to detect topics from entities, and subsequently incorporating logic to identify switches in a conversation, we observe abrupt topic changes in 4% of the user turns, such as (3):

(3) BOT: "Ok. So, i wanted to know, what's your favorite ice cream flavor?"
USER: "Let's talk about aliens."

In comparison, topic changes in human-to-human conversations generally occur in specific environments and in characterizable ways, and are rarely abrupt (Maynard, 1980).

Another Gricean maxim that appears not to apply in socialbot scenarios is the maxim of quantity, which requires responses to be appropriate in length. In interactions with the socialbot, however, user responses tend to be much briefer than one would expect in a human conversation, as in (4).

(4) BOT: "What do you think of the current state of the economy?"
USER: "Hit bad."

Overall the median utterance length for users is 3 words, much shorter than the bot's median utterance length of 21. In fact, almost 97.5% of user utterances are less than 14 words (see Appendix, Fig. 3). Such short responses are unusual in human conversations. This pattern might be due to the fact that users believe that the bot will be more likely to understand if they keep their responses short. Another possible explanation is that users feel it's the bot's job to drive the conversation forward, and thus take a more passive role.

The above examples make clear that many conventional conversational scripts don't apply to socialbot interactions. We find that many users employ bot-specific scripts, reverting to virtual assistant commands during conversations. Example (5) demonstrates a frequent phenomenon in the data: when a user feels the bot hasn't understood them, they invoke the standard prompts which they are accustomed to using when invoking the virtual assistant, by using the "Alexa" command to get the bot's attention and reset the prompts.

(5) USER: "Are you okay?"
BOT: "I am sorry I could not hear you well. Please repeat what you said."
USER: "Alexa, are you okay?"

38.9% of conversations include at least one invocation of the "Alexa" command. In these cases, instead of applying scripts from human conversations, users apply scripts they've learned from interacting with their virtual assistant. This tends to happen in cases where an unnatural or unsatisfactory response from the bot reminds the user that they are not chatting with a real human.

6 Discussion

One major difference between socialbot and typical human conversations is the perceived relationship between user and bot. In the user-socialbot relationship there is more of a power imbalance than in a human conversation; users are in control. They can stop, redirect, or reboot the bot, and choose conversation topics. The bot is designed to be cooperative, arguably more than a human when it comes to abrupt topic changes or overly brief responses. Where such responses might signal hostility (or at least disinterest) to a human interlocutor, users may consider such social implications irrelevant for a socialbot conversation.

Although all users are generally aware that they are speaking to a computer, some users are more willing to pretend. In the Alexa Prize, users were already users of the Alexa virtual assistant, and spoke to the socialbot on their Alexa-enabled devices. The socialbot uses the same voice as the virtual assistant, so the familiarity of the Alexa voice may foster a sense of the relationship between users and the socialbot, and allow some users to forget that they are interacting with a computer. Other users, however, will still be wary of human-like behaviors from the bot, as in (6).

(6) BOT: "I ate some pampered chef chicken salad tea sandwiches today, and it was amazing! Have you ever heard of it?"
USER: "No, Alexa. How can you eat something? You're a computer."

The Uncanny Valley is a clear obstacle to truly natural socialbot conversations, even if thresholds vary among users. Obviously, presenting a socialbot to a user as if it were really a human would pose ethical issues, so users' awareness of the conversation's artificiality is a necessary limitation.

7 Conclusion

The increasing quality and cultural salience of socialbots have led to significant advances in conversational AI. This paper analyzed conversations between an Alexa Prize socialbot and its users to better understand what users expect from socialbot interactions. We find that, because socialbots present a novel genre of conversation, users aren't always sure how to behave. Often, users react by applying human conversational norms to the socialbot; in other cases, they draw on the virtual assistant scripts acquired from using their Alexa-enabled devices. Based on our above analysis of user behavior, we feel that the goal of a socialbot shouldn't be to strictly mimic human conversation. Humans may be unpleasant, have diverging opinions, or push back on certain topics. On the other hand, socialbots are designed to provide an enjoyable and entertaining experience for the user. Socialbot developers should embrace the unique aspects of the scenario, rather than attempting to conform to conventional conversational norms.

We see two potential sources for the advancement of socialbot systems moving forward: first, developers should design bots to fulfill user expectations, acknowledging that these will be slightly different from human conversation norms. Second, as socialbots become more commonplace, the

emergence of socialbot-specific scripts will give users a clearer guide for those interactions. Like a real conversation, the future of socialbots must involve negotiating terms: developers must adapt socialbots to user expectations, and users will in turn adjust their expectations as they become more familiar with socialbots as a mode of interaction.

References

Ethan A. Chi, Chetanya Rastogi, Alexander Iyabor, Hari Sowrirajan, Avanika Narayan, and Ashwin Paranjape. 2021. Neural, neural everywhere: Controlled generation meets scaffolded, structured dialogue.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Sarah E Finch, James D Finch, Daniil Huryn, William Hutsell, Xiaoyuan Huang, Han He, and Jinho D Choi. 2021. An approach to inference-driven dialogue management within a social chatbot. *arXiv preprint arXiv:2111.00570*.

Kurt Gray and Daniel M Wegner. 2012. Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1):125–130.

H. P. Grice. 1989. *Studies in the Way of Words*. Cambridge: Harvard University Press.

Dell Hymes. 1972. Toward ethnographies of communication: The analysis of communicative events. *Language and social context*, pages 21–44.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, Ming Cheng, Qinglang Chen, Lauren Stubel, Karthik Gopalakrishnan, Kate Bland, Raefer Gabriel, Arindam Mandal, Dilek Hakkani-Tur, Gene Hwang, Nate Michel, Eric King, and Rohit Prasad. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize.

Jakub Konrád, Jan Pichl, Petr Marek, Petr Lorenc, Van Duy Ta, Ondřej Kobza, Lenka Hýlová, and Jan Šedivý. 2021. Alquist 4.0: Towards social intelligence using generative models and dialogue personalization.

Varvara Logacheva, Mikhail Burtsev, Valentin Malykh, Vadim Polulyakh, and Aleksandr Seliverstov. 2018. Convai dataset of topic-oriented human-to-chatbot dialogues. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 47–57. Springer.

Douglas W. Maynard. 1980. Placement of topic changes in conversation.

M. Mori. 1970. The uncanny valley. *Energy*, 7(4):33–35.

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

Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA. Association for Computing Machinery.

Marek N Posard and R Gordon Rinderknecht. 2015. Do people like working with computers more than human beings? *Computers in Human Behavior*, 51:232–238.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational ai: The science behind the alexa prize.

Sougata Saha, Souvik Das, Elizabeth Soper, Erin Pacquetet, and Rohini K. Srihari. 2021. Proto: A neural cocktail for generating appealing conversations.

Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of persuasive dialogues: testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Silvan Tomkins. 1987. Script theory. the emergence of personality. eds. joel arnoff, ai rabin, and robert a. zucker.

Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. 2021. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. *Developing a Personality Model for Speech-Based Conversational Agents Using the Psycholexical Approach*, page 1–14. Association for Computing Machinery, New York, NY, USA.

Marilyn Walker, Vrindavan Harrison, Juraj Juraska, Lena Reed, Kevin Bowden, Wen Cui, Omkar Patil, and Adwait Ratnaparkhi. 2021. Athena 2.0: Contextualized dialogue management for an Alexa Prize SocialBot. In *Proceedings of the 2021 Conference*

on Empirical Methods in Natural Language Processing: System Demonstrations, pages 124–133, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Terry Winograd, Fernando Flores, and Fernando F Flores. 1986. *Understanding computers and cognition: A new foundation for design*. Intellect Books.

A Appendix

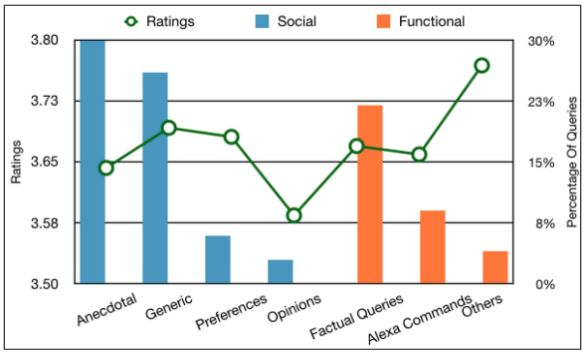

Figure 2: Analysis of different types of user queries. The primary Y-axis depicts the average rating associated with a query type across all conversations. The secondary Y-axis denotes the percentage of encountering each query.

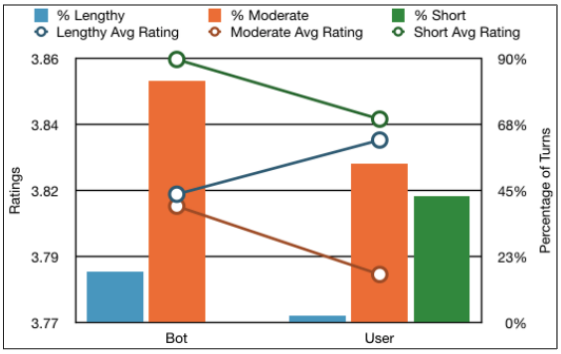

Figure 3: Analysis of bot and user response length. The primary Y-axis depicts the average rating associated with each length category across all conversations. The secondary Y-axis denotes the percentage of each length category for the bot and the user. Note that the percentage of short responses generated by the bot is very low.