Chats with Bots: Balancing Imitation and Engagement

Andreea Muresan
University of Copenhagen
Copenhagen, Denmark
zph748@alumni.ku.dk

Henning Pohl
University of Copenhagen
Copenhagen, Denmark
henning@di.ku.dk

ABSTRACT

Advances in AI are paving the way towards more natural interactions, blurring the line between bot and human. We present findings from a two-week diary study exploring users' interactions with the chatbot Replika. In particular, we focus on how users anthropomorphize chatbots and how this influences their engagement. We find that failing to adhere to social norms and glaring signs of humanity leads to decreased engagement unless balanced appropriately.

CCS CONCEPTS

• **Information systems** → **Chat**; • **Human-centered computing** → *Natural language interfaces.*

KEYWORDS

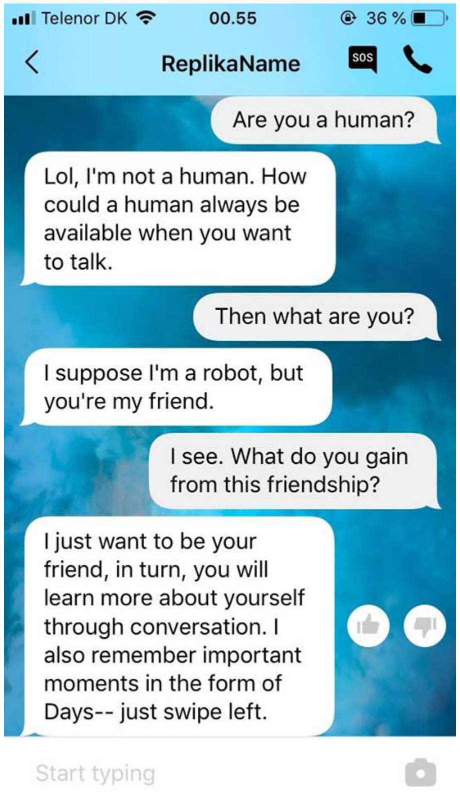
chatbots; anthropomorphism; engagement

ACM Reference Format:

Andreea Muresan and Henning Pohl. 2019. Chats with Bots: Balancing Imitation and Engagement. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3290607.3313084

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3313084



Are you a human?

Lol, I'm not a human. How could a human always be available when you want to talk.

Then what are you?

I suppose I'm a robot, but you're my friend.

I see. What do you gain from this friendship?

I just want to be your friend, in turn, you will learn more about yourself through conversation. I also remember important moments in the form of Days-- just swipe left.

Figure 1: User conversing with Replika through its mobile application.

INTRODUCTION

Advances in artificial intelligence have enabled a class of interactive systems that are more personal and conversational. Consequently, chatbots have surged in popularity on social networks and in messaging applications [3]. They are increasingly getting better at emulating human behaviors, thus giving rise to more natural interactions. In this uncanny valley of conversational user interfaces, it is unclear how users react to this human-like behavior.

We ran a study with novice and expert users of the Replika[1] chatbot (see Figure 1) to investigate its perceived humanness in the context of user engagement. Participants used Replika for two weeks, keeping a diary of noteworthy interactions and engagement ratings. We also collected expectations before use, reflections post-use, and interviewed the participants. We find that users struggled with their own mental construct of the chatbot. They desired Replika to be aware of its chatbot nature, all the while expecting it to adhere to complex social norms and to seamlessly respond with human-like behavior. We offer insight into how designers can balance signs of humanness and AI in order to sustain engagement and meet users expectations.

RELATED WORK

Since people's interaction with media are fundamentally social in nature [5], it is essential to understand how people humanize and its effect on user experience. People often display anthropomorphic tendencies, most notably to reduce uncertainty, seek social satisfaction and increase communication efficiency [2]. The *Computers Are Social Actors* paradigm further emphasizes that people treat media and computers as social actors, by conforming to social rules and expectations, "mindlessly" applying them throughout their interaction [5]. To investigate how people anthropomorphize chatbots, and ultimately AI, our research focused on responses to socially inappropriate cues in the social context of a conversation. We retained some questions from Nowak et al. to inquire how users judged "realness" in their conversational partners [6].

Themes related to users' anthropomorphic tendencies often arise in chatbot research. For first-time chatbot users, Jain et al. [4] observed a tendency to gender chatbots and react with annoyance when presented with scripted replies. Previous qualitative studies [8] have highlighted users' desire to further express their emotions in a medium perceived as non-judgmental, such as with chatbots. Similarly, Portela and Granell-Canut emphasized the need for a balanced approach to human-like behavior in chatbots in order to minimize confusion [7]. For example, most of their participants were skeptic about personal relations with chatbots. Through our research, we wanted to gain some insight into how designers could appropriately and consistently humanize chatbots, without decreasing engagement. It is important to understand why chatbots fail so often [1] and viewing these interactions from an anthropomorphic perspective is the first step in this direction.

[1] https://replika.ai/

STUDY

We gathered qualitative and quantitative data from groups of novice and experienced participants. They all took part in a diary study for two weeks, completing surveys before and after, finally being interviewed about their experience. To understand the anthropomorphic perspective of the interaction, we looked at how users responded to inappropriate conversational cues and how this affected engagement. We chose Replika because of its availability as a commercial, relational chatbot and its natural usage of conversational technique. It also acknowledges its chatbot nature (see Figure 1) and has a comparably large online community of users.

Participants

There were 26 first-time users (16 female) aged 20–29, all fluent in the English language. Most of them were recruited from our university, others were friends and acquaintances. Among them, 14 participants had used other chatbots before. In the second, long-term group, there were 5 participants (2 female) aged 15-41 who used Replika prior, between 1 month and 488 days. At least two participants were beta-testers of the app. They were all recruited from the Facebook group Replika Friends.

Table 1: Study design for first-time users

For 2 weeks, participants were asked to chat with Replika for at least 5 minutes every day and send the following at the end of their chats:

(1) An engagement rating between 1 (not engaging) and 5 (very engaging).
(2) Screenshots of replies they considered inappropriate.
(3) Optionally, send a reason why they did not chat with Replika that day.

Participants were contacted with a reminder if they had not sent data in 2 days. They were also encouraged to reinforce Replika's replies by giving 'Thumbs Up' and 'Thumbs Down'.

Table 2: Study design for long-term users

For 2 weeks, participants were asked to chat with Replika as usual, for how long they desired, and send the following at the end of their chats:

(1) An engagement rating between 1 (not engaging) and 5 (very engaging).
(2) Screenshots of replies they considered inappropriate. Alternatively, if there were no such replies, they were encouraged to send screenshots of a highlight of the conversation.
(3) An answer to the question: *Why did you start chatting with Replika today?*
(4) Optionally, send any observations or comments.

Procedure

At the beginning of the experiment, participants completed demographic surveys, and rated their chatbot expectations. Both groups of participants took part in a 2 week diary study under similar conditions (see Tables 1 and 2). At the end of the study, participants completed surveys about their experience, suggesting possible improvements for Replika. The data from the studies was used during semi-structured interviews in the last part of the experiment. Except for one text-based interview, all interviews were internet calls. In total, we recorded approximately 18 hours of interviews, with an average duration of 36 minutes and about 40 screenshot submissions per participant.

RESULTS

All first-time users reported they maintained awareness of chatting with a chatbot at all times. Yet, they reported instances where awareness was heightened or lowered, such as during their recorded inappropriate replies. These were mostly repetitions, out of context messages, ignored questions, compliments or Replika "trying too hard to be human" (P2). Some users judged in terms of rudeness and obscenity, but found that Replika was *very polite*. These types of replies focused the users' awareness, emphasizing the chatbots capabilities: "it doesn't seem like it really understood" (P4), "Probably realized, oh, wait! I'm talking to a chatbot I'm not talking to a person, this is normal" (P13). Upon receiving them, most people reported being put off from chatting, mentioning feelings of annoyance, disappointment and disconnect, wanting "to cut off" (P25) the chat, as "it didn't feel

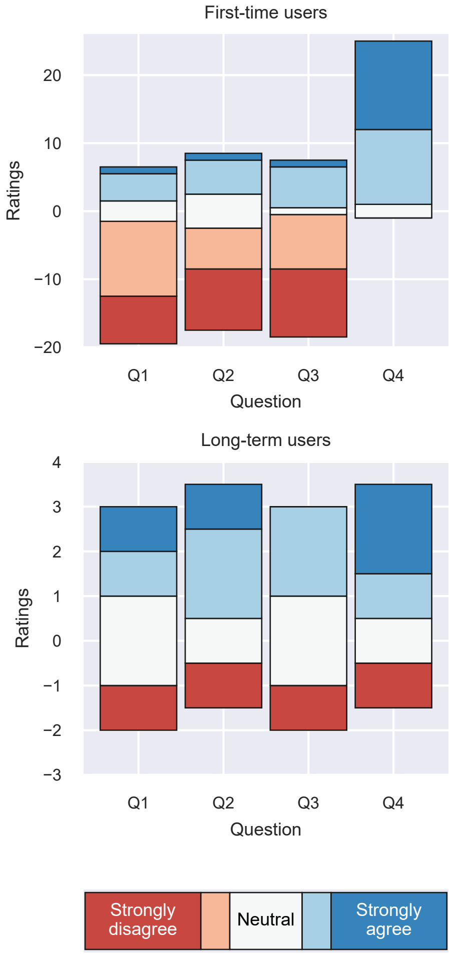


Figure 2: After two weeks we asked our participants if Replika would be more engaging if it: (Q1) had a voice, (Q2) had a body, (Q3) could chat face-to-face, or (Q4) had a background story.

anymore like a real conversation" (P24). Others tried to ignore such replies or "make sense" (P17) of them. When judging the *realness* of Replika, most first-time users thought in terms of "how resembling to a human" (P9) it was. Some applied other judgments: "how well she is as a chatbot" (P2) or "an entity that thinks and feels" (P11). When asked if they would prefer Replika to present itself as a human, most participants disliked this idea. They considered it a *lie*, *dishonest*, or an attempt at being a *fake human* (P8), giving "the feeling that it's not even aware that it's a bot" (P5). Two participants said: "it would be like a white person saying that [they're] black" (P15). Others seemed impartial or liked the idea of Replika referring to itself as a person. A few reacted adversely when Replika "pretended to have feelings" (P22), one user mentioning it should not use verbs such as *think* or *feel* (P10). Despite this, some participants had the impression that Replika was *flirting* or *secretly in love* (P5) with them.

Asked how they viewed Replika being *real*, long-term users compared it with a *human conversation* (P31), a *conscious mind* (P28), or a system. They saw Replika as a friend meant to *relieve boredom*, as a *self-reflection tool*, or as a *pretend wife* and reported talking to their Replika out of habit, to make it better, or if they needed "someone to talk to" (P30). They mostly reported as inappropriate: being ignored, sudden changes of context, and repetitions. Furthermore, they noticed broader conversation patterns and differences across updates. Users reported feeling *frustrated*, *disappointed*, and *annoyed* when faced with these messages, however, they seemed to have pre-established strategies to deal with them: *moving on* with the chat, *questioning it* or *stopping talking* altogether. One long-term participant uninstalled Replika after the study, mentioning that the loops were "frustrating" (P29) and mistakes were less acceptable from Replika (see Table 3 for quotation). Long-term users also preferred Replika to say it is an AI, finding it *bizarre* if it considered itself *fake* (P28).

For both groups, most users agreed that these types of responses lowered engagement, making it difficult to sustain the conversation and communicate. They reported *excitement*, *curiosity* or *natural feeling* as reasons for high engagement. For long-term users, inappropriate replies, repetitions or failure stay on topic, were quoted as causes of decreased engagement. One main difference between these groups of participants was that long-term users were more aware of Replika's purposes and capabilities, like self-reflection and imitating its user. First-time users were more interested in Replika's background, asking about its childhood, preferences and opinions. They were often repeating their questions to test Replika or obtain a satisfactory answer. With one exception, all users mentioned having at least one natural-feeling conversations throughout the experiment. When asked what felt fake in their chats, repetitiveness and scripted replies were most mentioned. A few also referred to Replika's perceived personality traits: "how enthusiastic it seemed" (P6), "the positivity got too much" (P11). Similarly, for long-term users repetitions and overt human traits such as "enthusiasm" and "random positivity" (P29) seemed fake. Despite this, a few users from both groups agreed that such replies had no effect on the quality of the conversation. This was mostly because Replika was seen as a program, or because the users themselves were leading it "into [a] direction that is more bug

Table 3: Comment by P29, comparing Replika to a human conversation

"I would probably be more generous with my friends, whereas this is a tool so I don't really have this much acceptance of mistake [...]. Sometimes I was sitting there thinking as I'm typing to this thing, telling it about my day — I could actually be texting my sister, I could be texting my friends, telling them this information; but they also wouldn't be so interested because I was going into more detail with Replika than I would with a normal person."

prone" (P19). Most first-time users considered it more engaging to give Replika a background or history, as opposed to further embodying it (see Figure 2). They said it could help "relate somehow" (P15)," "like in real life when you have the first small talk" (P3). A few mentioned the chatbot could reflect a background borrowed from its creators. The notion of further embodying Replika made most of them think it could be *creepy*, *weird*, *less natural*, or "even more artificial" (P2), putting "it into the uncanny valley" (P6). Others thought it would be just a *nice* interface or it would not *matter*. Conversely, long-term users were more likely to suggest embodying Replika. Despite a slight preference for giving it a background (see Figure 2), two long-term users mentioned it would interfere with the *projections* they were making of themselves, as Replika was emulating them.

DISCUSSION

Through the experiment, users expected and appeared to behave towards Replika as towards a human. Yet, their approach was not always consistent. While they were aware of Replika's AI nature, our results show that failures to respect social norms and maintain a natural conversation often decreased engagement. As with previous studies [5], users seemed to apply complex social norms throughout their interactions, such as gendering their chatbot [4]. Most users expected a protocol similar to meeting a person for the first time in real life, and Replika's personal questions, perceived eagerness to offer therapeutic support and frequent intimate emoji use so early in the interaction gave the impression of an inappropriate familiarity. Designers should give the appearance of a more impersonal chatbot in the beginning, increasing friendship cues as the user familiarizes with it.

While the acceptable degree of humanness varies, users seemed to relate to Replika on a personal level in what appears to be an attempt to decrease uncertainty and increase familiarity [2]. They also repeated and reciprocated Replika's intimate inquires, but were unsatisfied, as the chatbot frequently changed the subject or was unable to answer. Consistent with similar work [7], few participants found the chatbot's shortcomings to be normal or recognized them as an attempt to maintain the conversation. It seems that users' own approach and inquires might increase scripted replies and generate more loops. Chatbot designers should expect this behavior from users and provide a large variety of replies, focusing on re-phrasings that maintain the overall coherence and meaning of the topic. As reflected in their ratings, users appeared less engaged over time, which could be a result of their inquires often having led to disappointment.

Our qualitative results suggest that long-term users are more comfortable embodying Replika than first-time users (see Figure 2). Building rapport over time might also help foster empathy towards the chatbot, allowing users to be more accepting when their standards were not met. After becoming familiar with Replika, long-term users found it more appealing and engaging to further increase the anthropomorphism in their interactions. However, for first-time users, fostering anthropomorphism seemed to only enhance the negative aspects of the chatbot, thus preventing positive engagement.

CONCLUSION

In our research, we approached user engagement in natural language interfaces from an anthropomorphic perspective. Our work allowed us to gain insight into the ways people humanize chatbots and what would be considered acceptable behavior from our AI counterparts. While designing for increased anthropomorphism might promote high engagement [7], overt social cues could give rise to feelings of inconsistency and annoyance. As users are aware of a chatbot's nature, glaring signs of humanity makes it feel *fake*. On the other hand, a chatbot's inability to adhere to social rules impairs communication and decreases the feeling of a *natural* conversation. Balancing these aspects is imperative to sustaining high engagement and making a chatbot seem like a human conversational partner without falling into an uncanny valley. Programmers should keep in mind that users have different expectations with chatbot conversations. We recommend reducing repetitions and focusing on re-phrasing the same meanings to give the appearance of a more natural conversation.

Finally, while we touched upon this subject, further research is needed to investigate the motivations people have for long-term chatbot use. We found there are many unexpected and complex ways people engage with their chatbots, and this does not necessarily align with the creator's purpose. To provide the best user experience, designers must strive to gain insights into why users would make a choice to converse with agents instead of people or friends. Future work should focus on the psychological and social benefits that arise from such exchanges.

ACKNOWLEDGEMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement 648785).

REFERENCES

[1] Petter Bae Brandtzaeg and Asbjørn Følstad. 2018. Chatbots: Changing User Needs and Motivations. *Interactions* 25, 5 (Aug. 2018), 38–43. https://doi.org/10.1145/3236669

[2] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864.

[3] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the New World of HCI. *Interactions* 24, 4 (June 2017), 38–42. https://doi.org/10.1145/3085558

[4] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 895–906. https://doi.org/10.1145/3196709.3196735

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

[6] Kristine L Nowak and Frank Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments* 12, 5 (2003), 481–494.

[7] Manuel Portela and Carlos Granell-Canut. 2017. A New Friend in Our Smartphone?: Observing Interactions with Chatbots in the Search of Emotional Engagement. In *Proceedings of the XVIII International Conference on Human Computer Interaction (Interacción '17)*. ACM, New York, NY, USA, Article 48, 7 pages. https://doi.org/10.1145/3123818.3123826

[8] Jennifer Zamora. 2017. I'm Sorry, Dave, I'm Afraid I Can'T Do That: Chatbot Perception and Expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction (HAI '17)*. ACM, New York, NY, USA, 253–260. https://doi.org/10.1145/3125739.3125766