AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support

Jess Hohenstein
Malte Jung
Department of Information Science
Cornell University
Ithaca, NY, USA 14853
jch378@cornell.edu
mfj28@cornell.edu

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada

ACM ISBN 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3188487.

Abstract
Despite a growing body of research about the design and use of conversational agents, existing work has almost exclusively focused on interactions between an agent and a human. Less is known about how an agent is perceived and used during human-human conversation. We compared conversations between dyads using AI-assisted and standard messaging apps and elicited qualitative feedback from users of the AI-assisted messaging app through interviews. We find discrepancies between the AI assistant's suggestions and the conversational content, which is also reflected in participant interviews. Our results are used to suggest some areas for improvement and future work in AI-assisted communication.

Author Keywords
Artificial intelligence; CMC; Google Allo; messaging.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

Introduction
From customer service, social and emotional support, and entertainment, conversational agents are

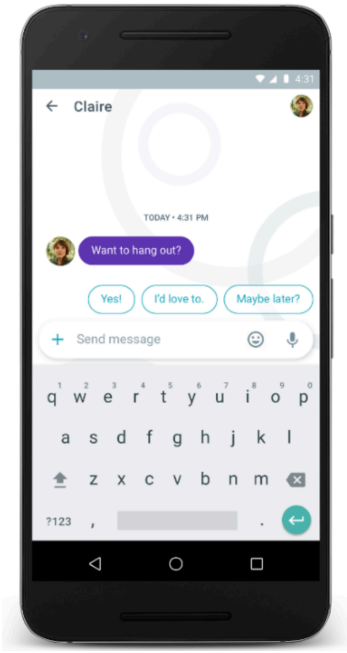

Figure 1: Google Allo is an AI-assisted messaging app that includes suggested responses that the user can tap on to quickly reply, as shown above.

increasingly becoming a part of people's daily lives. In addition to interacting directly with individuals, chatbots are also beginning to participate in interactions between people. As AI-assisted communication becomes more prevalent, it is increasingly important to understand how to make such technology as relevant and useable as possible, as well as the effects that it may have on human-human interaction.

Despite the growing body of research about the design and use of chatbots, little is known about how an agent is perceived or interacted with during AI-assisted conversation between humans. We investigate this by comparing conversations with a traditional messaging app and an AI-assisted messaging app. We find discrepancies between the AI assistant's suggestions and the conversational content, which may have resulted in the low usage of the suggestions that was quantitatively observed.

Background

Advances in AI and machine learning combined with increased adoption of mobile messaging platforms have fueled a growing interest in chatbots [3]. To improve these experiences, research has investigated how people interact with conversational agents and found that chatbots are mainly used for productivity, and, to a lesser degree, entertainment, social factors, and novelty [1]. Developers have in turn attempted to meet user requirements by making chatbots with flexible capabilities, unambiguous messages, and reliable user interactions [7]. Despite the potential benefits and uses of embodied conversational agents, overall user adoption has been less extensive than originally envisioned [10], which could be explained by agents' noted shortcomings, including an inability to meet expectations due to lack of understanding and insufficient usability [2] or an inability to engage emotionally [1].

Building on the extensive work in human-chatbot interaction, some systems have investigated the ways that human-human messaging could be improved with AI assistants, [e.g. 4,6]. While these and similar systems have shown some benefits of AI assistants on human-human interaction, the prototypes used were not robust and the conversational contexts were very specific, limiting the generalizability of such findings. Despite this work and the well-established promises and difficulties of human-chatbot messaging, there is a lack of studies that investigate the use and effects of AI agents on human-human conversation. Our investigation addresses this gap by investigating how users did (or not did) interact with an agent during a messaging conversation and why.

Method

To investigate the use of an AI assistant in human-human conversation, we asked participants to carry out a communication task with either a traditional messaging app or with an AI-assisted messaging app.

Google Allo

Google Allo combines an AI assistant with text messaging to create an AI-assisted messaging app and was the focus of this study. Users can invoke a Google web search within the app, and the app also frequently provides "Smart Replies", suggested responses based on an algorithm and parsing of the conversation history. The suggested responses typically come in groups of three after a message is sent or received, as shown in Figure 1. To our knowledge, this is the first

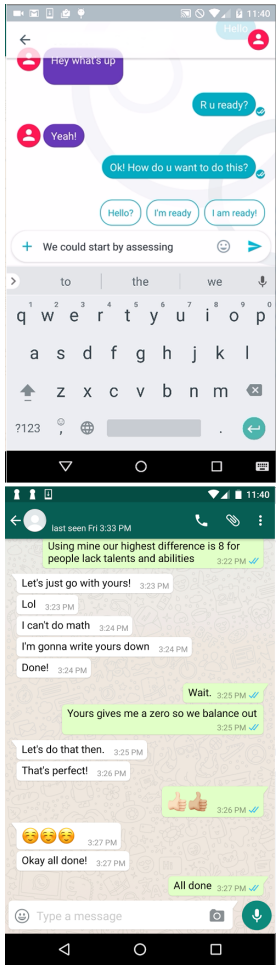

Figure 2: Google Allo (top) and Whatsapp (bottom) were the messaging apps examined in this study.

app of its kind to use an AI assistant to mediate text messaging conversations between two individuals.

Participants and Tasks

Participants (N=72, 36 dyads) were recruited through an on-campus recruiting system at a large university in the United States and received either course credit or monetary payment for participating. Ages ranged from 18 to 27 (M=21.0, SD=1.82). In terms of gender, 50 participants identified as female, and 22 participants identified as male. Racially, respondents identified as 37.5% East Asian, 31.3% Caucasian, 10.4% South Asian, 10.4% Hispanic/Latino/Chicano/Puerto Rican, 4.2% Black Non-American, 2.1% Black American, 2.1% Bi-racial/Mixed/Multi-racial, and 2.1% Pacific Islander.

Participants used either Google Allo or a standard messaging app (Whatsapp) to complete one of three tasks. These tasks included the Causes of Poverty task [8], where the pair was asked to come up with an agreed-upon ranking of 13 causes of poverty, the Lifeboat task, where the pair was asked to come up with an agreed-upon ranking of 5 people who should get a spot on a lifeboat, and the Trip Planning task, where each participant in the pair was given a different budget and asked to plan a trip for the upcoming weekend. In each task, participants were motivated to negotiate through the promise of extra compensation based on how close the final ranking/budget was to their initial ranking/budget. Participants were given a time limit of 1 hour, which we assumed, based on informal pilot testing, would be plenty of time for most to finish any of the tasks.

Equipment and Recording Processing

Participants used Android Nexus 6 smartphones running Android 7.0. Whatsapp was chosen as the control messaging app because of its similarities in UI to Google Allo, as shown in Figure 2. Screen recordings were taken of each conversation using the AZ Screen Recorder app for later transcription and review.

Procedure

The study consisted of two parts, the experiment and the qualitative interviews, with only selected participants participating in the latter. For the experiment, two participants were present for each trial. Each participant was placed in a separate room and remained completely anonymous to the other. If participants were going to use Google Allo, they were asked to review a sheet of information about its noteworthy features. Next, participants read about the task and the scoring procedure, reviewed the app, and asked any questions. When ready, one participant informed the researcher that they were going to begin the conversation and sent the first message to their partner. Conversation histories were cleared between each pair of participants.

Five random participants from the Allo group were asked to return for interviews. While viewing a screen recording of the conversation, they were asked to "think aloud", elaborate on their thoughts in retrospect, and comment on the app itself and the suggested responses. After viewing and commenting on their conversation, participants were asked some additional questions about using Allo. These sessions were recorded and transcribed.

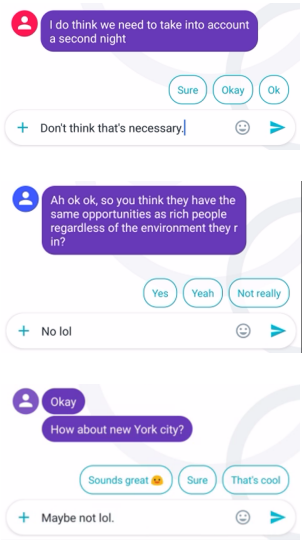

Figure 3: Allo's suggested responses are not relevant to what the user wants to express.

Results and Discussion

For the analysis, we examined 36 conversations, including 18 with AI-assisted messaging via Allo and 18 with standard messaging via Whatsapp.

Upon reviewing the conversations, one of the first things that we noticed was the overwhelming positive sentiment of the suggested responses. Suggested responses with a positive sentiment (e.g. "Yes", "Sure", "Nice", "Right", "Ok", etc.) seemed to be noticeably more common than suggested responses with a negative sentiment (e.g. "Me neither", "Not really", "No", "Nothing", etc.). Upon classifying the sentiment of the suggested responses using Mechanical Turk (5 workers per response; 1012 unique suggested responses), we found that 43.8% of all suggested responses were classified as having a positive sentiment, while only 3.95% were classified as having a negative sentiment. This finding led us to wonder whether this excess of positive compared to negative suggested responses was having a priming effect on conversational dynamics or affecting outcomes.

Among the participants that conversed using Google Allo, the suggested responses were hardly ever used. On average, among the 36 participants who used Allo, a suggested response was chosen 6.24% of the time. Each instance of a response suggestion was defined as any time Allo displayed a selection of suggested responses, regardless of quantity, to the user. This finding may be linked to the nature of the tasks coupled with the seemingly positive skew of the suggested responses. Upon reviewing the conversations, we also noticed that, especially in moments of contention, Allo did not suggest responses that reflected what participants actually wanted to express. Examples of this can be seen in Figure 3, where users wanted to disagree or suggest an opposing viewpoint, but Allo's suggestions only offered expressions of agreement and positive reinforcement.

Interviews

Qualitative results from participant interviews generally reflect the quantitative findings. Participants expressed that while they could see the potential usability of the suggested responses, they were not particularly relevant to the completion of the task. Participant 16 said that, "Since the task was pretty specific, maybe not all suggestions were exactly useful, but there were some that were nice. In terms of general contact between a friend, I think that it would be helpful in terms of, 'Hey how are you doing', and stuff like that." Similarly, Participant 12 told us that if she were having a "more casual conversation", the suggested responses would be helpful "so I don't have to physically type it out while I'm trying to cross the street, just press a button." Despite not using the suggested responses frequently, some participants also expressed surprise regarding their relevance to the conversation. Participant 16 said, "It was surprising to see how applicable they were to what I was saying in the context of our conversation."

Participants also honed in on why the suggested responses were largely unused. Participant 28 explained her desire to type in her own words: "Even when I'm text messaging, I like to type everything out...I like to type out the entire word. In general when I'm texting, I'm very particular". Other participants discussed how they simply did not notice the suggested responses at the time, with Participant 8 saying, "One of the options was 'Oh wait'; I should have just pressed

it… I just didn't notice it". Participants also discussed how the responses were often not relevant to their specific topic of conversation. Upon seeing some of the suggested responses, Participant 10 reflected, "Sometimes I was like 'What? This is so out of context'", and "I wished that there were suggested responses that made sense to use… just that what was suggested made more sense".

Some participants also speculated about the possible influence of the suggested responses on their conversation. Participant 8 explained how the suggestions "…would kind of guide me. It was what I was already thinking in my head, but then like okay, that's also an appropriate thing to say". Participant 28 explained, "Let's say if the other person asked for specific rank, and the prompts are all positive, and then I just agreed with it, I just went with the positive one just for the sake of completing the task. But if I hadn't seen the prompts, maybe I would have opposed that question." Participant 12 told us about how, regardless of their relevance to the current conversation, "it was very tempting to click the emojis sometimes".

Design Suggestions

The infrequency with which Allo's suggested responses were used by participants could be due to the mismatch between the suggested responses and the conversational content. Users might be more likely to choose the suggested responses if they were more relevant to the tone of the conversation. For example, Figure 3 shows some of the numerous examples of a user wanting to disagree with their partner's statement but only being offered positive suggested responses by Allo. It could be beneficial if, each time responses were suggested, users were presented with a mix of positive and negative options. Additionally, since Google already acquires data from conversations held using Allo [4], the sentiment of these suggestions could quantitatively correspond to the negative and positive makeup of conversations in the database. This could be accomplished through amending the reply-generating beam search to be biased towards paths leading to responses that reflect this sentiment makeup.

Many current messaging apps offer users an option to "autofill", where words are dynamically suggested based on what the user has already typed. A similar feature implemented in Google Allo, where suggested responses are generated as the user is typing, could potentially make suggested responses more relevant to users. Participant 28 mentioned, "Let's say I'm typing 'N'; if the prompts are related to that specific letter they would be more helpful". Taking the last instance in Figure 3 as an example, when the user started typing "M", Allo could have generated suggested responses of "Maybe", "Maybe not", and "Maybe later", which would have been more relevant to what the user wanted to express.

Limitations

There were several limitations to this study. First, we analyzed conversations from participants completing a negotiation task. This type of conversation was likely not representative of an everyday messaging conversation, and the results found here may not generalize to other types of conversations. Future work should examine the effect, if any, of AI-assisted messaging on conversations with various content.

Our sample consisted of university students in the United States, and the results might not generalize to

other populations. Research has shown that people from various cultures may have vastly different impressions of the same AI assistant behavior [4]. Future work would benefit from examining the use and acceptability of AI-assisted communication among people of different cultures and ages. However, young adults are the most avid users of text messaging by a wide margin [9], so this sample is useful for understanding everyday perceptions of AI-assisted messaging.

Conclusion

As AI assistants continue to become increasingly prevalent, it is important to understand how they are used in interactions between humans. Despite a growing body of research regarding AI assistants, existing work has mostly focused on interactions between humans and AI. To our knowledge, this is the first study investigating the effects and use of an AI agent on messaging conversations between humans.

We found that the AI suggested responses were infrequently used and that discrepancies existed between the AI suggestions and the conversational content. These findings were mirrored in the qualitative feedback from participants, as users elaborated on how the responses were inappropriate or not relevant to the conversation. Additionally, participants indicated that the AI-assisted messaging app may have influenced what and how they communicated.

References

1. Petter Bae Brandtzæg and Asbjørn Følstad. 2017. "Why People Use Chatbots." In *International Conference on Internet Science*, 377-392.
2. David Coniam. 2014. "The linguistic accuracy of chatbots: usability from an ESL perspective." *Text & Talk* 34(5), 545-567.
3. Asbjørn Følstad and Petter Bae Brandtzæg. 2017. "Chatbots and the new world of HCI." *interactions* 24(4), 38-42.
4. Katherine Ibister, Hideyuki Nakanishi, Toru Ishida, and Cliff Nass. 2000. "Helper Agent: Designing an Assistant for Human-Human Interaction in a Virtual Meeting Space". *CHI Letters* 2(1), 57-64.
5. Nathan Ingraham. 2016. "Google stores 'transient' Allo messages until you delete them". Engadget. https://www.engadget.com/2016/09/21/google-allo-messages-privacy/
6. Joseph Kim and Julie A. Shah. 2016. "Improving Team's Consistency of Understanding in Meetings". *IEEE Transactions on Human-Machine Systems* 46(5), 625-637.
7. Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. "The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms." In *Proceedings of the 2017 Conference on Designing Interactive Systems*, 555-565.
8. Daniel T. L. Shek. 2002. "Chinese adolescents' explanations of poverty: The perceived causes of poverty scale". *Adolescence* 37(148), 789-803.
9. Aaron Smith. 2011. "Americans and Text Messaging". Pew Research Center. http://www.pewinternet.org/2011/09/19/americans-and-text-messaging/
10. Tom Simonite. 2017. "Facebook's Perfect, Impossible Chatbot". MIT Technology Review. https://www.technologyreview.com/s/604117/facebooks-perfect-impossible-chatbot.