An Exploratory Study on Conversational Agents Using Dynamic Conversation Styles

Bar Lerer[1] , **Sihang Qui**[1] , **Ujwal Gadiraju**[1] , **Jie Yang**[1]
[1]TU Delft

Abstract

This study aims to examine the effect of dynamically aligning the conversation style to the user's preference in a conversational agent environment. We suggest a technique to identify the user's preferred conversation style, by scoring statements based on their writing style. We designed a within-subject experiment intended to measure the user engagement and satisfaction (with style alignment × without style alignment). We found an increase in self-reported satisfaction ratings and reflect on how familiarity with the entity behind the chatbot may affect the preference of the conversational style. We finish this study with concrete suggestions to designers and developers of future chatbots and how should the alignment of conversation style be assigned and adapted.

1 Introduction

Conversational agents (CAs) have seen a steady increase in popularity over recent years [1]. While CAs have been implemented on many platforms, their objective is similar, to emulate as much as possible the "human" side of a conversation and to follow the flow which resembles texting with a friend or a relative [2]. Kiesler et al. identified an issue with unfamiliar context when conversing with a computer. The unpredictable style of messages might make the conversation difficult to understand [3]. There is a big challenge in developing CAs, as there is a gap in how the user wants to interact with the system, and what is anticipated from the user to say [4]. In a paper written by Tannen [5], she writes "'Style' is not something extra, added on like frosting on a cake. It is the stuff of which the linguistic cake is made: pitch, amplitude, intonation, voice quality, lexical and syntactic choice, rate of speech and turn-taking, as well as what is said and how discourse cohesion is achieved". The conversation style employed by a CA has an important role in ensuring the conversation is pleasant to the user, as there exists the risk that the user simply will not engage with the CA [6].

While the application domain is growing for CAs, in this paper, the focus would highlight the usage of chatbots for survey answering. The term 'chatbot' is used here to refer to a textual representation of a CA [7]. Under certain conditions, chatbots can perform better than the standard survey practices employed today. A research experiment found that when CAs were combined with an informal language, which resembled a natural conversation with a human being, produced higher quality data, compared to a standard web survey [8]. The authors compared a chatbot against a standard HTML survey, using both a formal and informal conversation style. Surprisingly, the experiment also concluded that there was no difference in the quality of the responses between a chatbot and an HTML survey when a formal language was used. These results imply that conversation style is one of the factors that influence the interaction with the CA.

Conversational style contains many linguistic elements [9, 10]. In this paper, we will focus on the formality of the conversation. Informal conversation style, can be defined as "common, non-official, familiar, casual, and often colloquial, and contrasts in these senses with formal" [11]. It may include, for example, the usage of emojis, text abbreviations, and humor in the conversation. A more static conversation, which is neutral in tone and lacks in personality can be seen to match a formal conversation style.

The results produced by [8] fall in with statistical findings of a study on conversation agents usage, where 25% of users would avoid talking with a CA because "it was not able to chat in a friendly manner" [6]. The authors acknowledged, based on Role Theory, that they were expecting some users to find the informal conversation style to be inappropriate. While their experiment did not reach the same conclusion, it is within reason to assume some users will find the informal style to be inappropriate. They concluded that companies that make use of CAs should research their user group and try to better understand what they would deem as appropriate. As CAs are becoming more popular, it is to be expected that more companies and organizations will make use of them [6]. However, having each company research its user base (which can change

over time) for the sole purpose of understanding which conversation style is appropriate, can become a lengthy and costly process.

This research aims to fill the gap on finding the appropriate conversation style, by suggesting an approach to determine the user's preferred conversation style and align the chatbot's conversation style based on the preference. This study aims to answer the following research question:

RQ. How can personalized conversation styles improve the satisfaction and engagement of users?

Building on the assumptions made by Liebrecht et al. [6] where some users would find informal style inappropriate, and the research results of [8], the hypotheses to be researched are as follows:

Hypothesis 1 *Using dynamically aligned conversation style to the user's preference will lead to higher engagement, measured using the UES-SF scale [12].*

Hypothesis 2 *Using dynamically aligned conversation style to the user's preference will lead to higher user satisfaction, compared to a conversation style that was not aligned.*

We created a pre-task survey as a heuristic technique to estimate the user's preferred conversation style. We proceeded to create a within subject experiment (with style alignment $\times$ without style alignment), testing the user's engagement and satisfaction ratings when the conversation style was aligned to his preference. We have measured the engagement of the user using the UES-SF [12]. Satisfaction was measured with a 5 point scale for comparison between the different environments. The users have answered a general survey to demonstrate the interaction with the different styles. Our results show that using conversation style alignment can increase user satisfaction among participants. We found that conversation style alignment had little to no influence on user engagement. Our work outlines design implications based on the results to future creators of conversational agents.

2 Related Work

2.1 Effects of Conversational Style

There is a large number of published studies [8, 6, 11] that describe the effect of a chatbot's conversational style on how the interaction was perceived by the user. The study of Kim et al. [8] conducted a 2 (platform: chatbot and web-survey)$\times$2(conversations style: formal and informal) research experiment, where they compared the responses received by the users in each of the environments. They have found that the interactivity in a chatbot environment produces higher-quality data, but when accompanied with an appropriate style. It is implied that an informal conversational style, which resembles a human-to-human interaction would be more suited to be used by a chatbot.

The work of [6] studied the effect of brand familiarity on the desired conversation style. Through a series of experiments, they have explored the changes in users' trust towards a brand, depending on the conversation style that was used. Their research concluded that the usage of informal style in the context of social media, when the brand was known to the users has increased the trust in the brand. The opposite effect was found when an informal conversation style was employed by a brand that was unknown to the users. They found it to decrease trust in the brand. In their final remarks, they have stated that the finding opposes previous research which found the informal style to be more suited in a social context, and that brand familiarity is something that should be addressed when deciding on an appropriate conversation style.

2.2 Conversational Agents as Crowdsourcing Platforms

Previous works have utilized chatbots as a crowdsourcing platform [13, 14, 15], arguing they can have better worker output than the standard crowdsourcing platforms, such as Amazon Mechanical Turk (AMT).

Previous studies have examined whether a conversational agent can serve as a proper alternative to crowdsourcing platforms [7, 16]. To better understand whether a CA can be used as a crowdsourcing platform, Mavridis et al. [7] have examined different types of crowdsourcing tasks such as CAPTCHA answering, or questions with multiple-choice answers, and have proposed conversational interface alternatives to these types of questions. Their results showed that a conversational agent can be used for crowdsourcing tasks, as the output and execution time was comparable between the different platforms.

A large experiment including 800 participants has examined the effect of both the conversation style and the platform on which a crowdsourcing task is performed [13]. They have found that using a conversational agent rather than the alternatives can have benefits in worker engagement and the usage of the appropriate conversation style can improve worker retention.

2.3 User Engagement And Satisfaction

Defining engagement and satisfaction is an important step to be able to measure improvement using existing scales. We use the definition of engagement from the works of [12] as "User engagement is a quality of user experience characterized by the depth of an actor's investment when interacting with a digital system". To measure user engagement across different platforms or experiences, the User-Engagement-Scale (UES) was developed. The UES is composed of 31 questions, spanned across six factors. An empirical evaluation of the UES was performed across three studies found the UES to be a reliable and valid scale in measuring user engagement, in the context of the online news domain [17]. A paper which was published in 2018 criticizes the length of the UES questionnaire and found that a

long-term evaluation of the scale is difficult to conduct, as many researchers have omitted certain parts of the UES [12]. For this reason, the authors have formulated the User Engagement Scale Short Form (UES-SF), a 12 question survey that was determined to be as reliable as the original UES, while offering a shorter experience for the users. The paper also includes an instruction manual for researchers on how to use the UES-SF. The instructions were implemented in this research.

For satisfaction, we define it as used in the paper [18] "as an emotional response or affect toward an object." To the best of our ability, no satisfaction measuring tool was found in the context of interactions with computers. For this reason, we have decided to use a simple five-point scale to allow the users to rate their interaction.

3 Methodology

In this section, we will present the method that will guide the decision to use either a formal or informal conversation style in a conversational agent environment when interacting with the user. After which the system that was used to collect the responses from the users would be introduced and the section would end in the description of the questions that were used to converse with the users.

3.1 Determining the User Preferred Conversation Style

We suggest the following method to identify the user preferred conversation style, which is based on an approach made to estimate the user's conversation style [19]. We asked a series of pre-task questions, asking the user to rate on a scale from 1 to 5, how do they like the style of writing for four responses, which are similar in meaning but differ in formality. The responses were formulated based on attributes from informal and formal conversation writing styles. These responses can be seen in Table 1.

Once the scoring of each statement was saved, the sum of both statements representing a conversation style was taken, and the higher sum pair would determine the preference of the user. If the combined ranking of questions 1 and 3 was higher than the combined ranking of questions 2 and 4, then the preferred conversation style was set to be informal style, and otherwise to a formal style. When the ranking was identical for the two pairs, the user was assigned with no preference and received a random styled survey.

To rule out any influence caused by differences in the question description between the different styles, we have only adapted the responses to the users' input, and the instructions for answering every question. An example of the difference in the conversation styles can be seen in Figure 1.

3.2 Dandelion System

To interact with the users, the Dandelion system [20] was chosen. Dandelion is a crowd computing platform

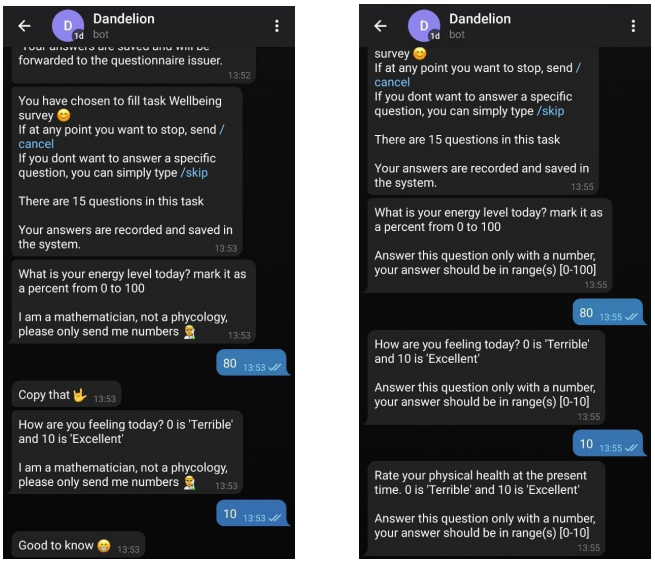

Figure 1: Comparison of informal (left) to a formal (right) conversation style, using the Dandelion system. Only the responses and instructions were adapted.

allowing for interaction between students and university researchers at TU Delft. The system allows the research team, to send tasks in the form of surveys to any participant(s) who registered to the system.

The Dandelion system lacked support for sending responses and instructions based on the preferred style. This functionality was added to the conversation creator component of the system. The added functionality was developed in a scalable way to allow for more conversation styles in the future. The system offers analysis tools, which allowed for easy classification of the preferred style to each user. Furthermore, the system is hosted on TU Delft infrastructure, and it allowed to send scheduled tasks in the form of a survey, which is why it was chosen for this research.

3.3 Survey

The questions that composed the task were adapted from the "Well Being survey" sent by TU Delft [21]. There are a total of 13 questions. The questions tried to determine the user's general well-being and how it was affected in the times of the recent COVID-19 pandemic. The questions are general and do not make any assumptions of the user, which is why it has allowed for more participants to partake in the experiment. Before answering any questions, the participants were introduced with instructions on what this research aims to answer and were instructed to focus on comparing between the different conversation styles, rather than focusing on providing complete answers to the survey.

3.4 Procedure and Measures

This research employed a within-subjects design, where every user would try both conversation styles and would report on the experience. The entire experiment was

Question Number	Text	Writing style
1	"I love going to the beach 🏝️🏊, but honestly I hate running there as is always sand in my shoes!!!"	Informal
2	"I enjoy the beach, but running there is quite irritating because of the sand"	Formal
3	"I swear, these Computer Science 💻 lecture are difficult and long! but they are interesting."	Informal
4	"While the lectures can have increased difficulty than the norm, they do contain interesting aspects"	Formal

Table 1: Questions to identify the preferred conversation style

estimated to take 20 minutes. To mitigate any potential influence caused by a learning bias of the survey, the setup was designed as two experiments, where the first group have answered a pre-task survey (See Figure 2 for example), followed by the well-being survey based on their preferred style. Upon completion of the survey, a second survey was sent, with the lesser preferred conversation style. A second group would follow the same steps, but the order of the styles was reversed, meaning they first answered a survey with the lesser preferred style, followed by the preferred style. After each survey, the engagement of the users was evaluated using the User Engagement Scale Short Form (UES-SF). In addition to the UES-SF, the users were also asked to rate their satisfaction from the interaction on a scale from 1 to 5, stating how much they have enjoyed the conversation. Using both engagement and satisfaction scales allowed to recognize whether the preferred style had increased satisfaction and engagement. Seeing that the UES would have been evaluated twice for every user, it was decided the use the shorter form (UES-SF) [12], which is congruous with the UES.

3.5 Participants

A total of 42 participants attempted the experiment, which were recruited using the Prolific platform [22]. 30 participants have completed the experiment ($M_{age} = 25, SD_{age} = 7.06$). The average time to complete the experiment was 14.8 minutes. The participants were paid for their relative participation time on the experiment.

4 Results

As can be seen in Figure 3, 16 participants preferred formal conversation style, 9 preferred informal and 5 participants had no preference towards any of the styles. As the sample size in this experiment was relatively small, no significance testing was performed.

4.1 Satisfaction

After conducting each survey using a different style, the participants were asked to rate on a 5-point scale how satisfied were they with the interaction. The goal was to measure the change in satisfaction levels between the different conversation alignments. From the following results, the group of participants without a preference towards a specific style were excluded. Their assesment would be discussed in section 4.3. A total of 9 people, out of 25 with conversation style preference, reported an increase in satisfaction. The increased satisfaction scores, have $M = 1.4, SD = 0.72, N = 9$. Overall, when

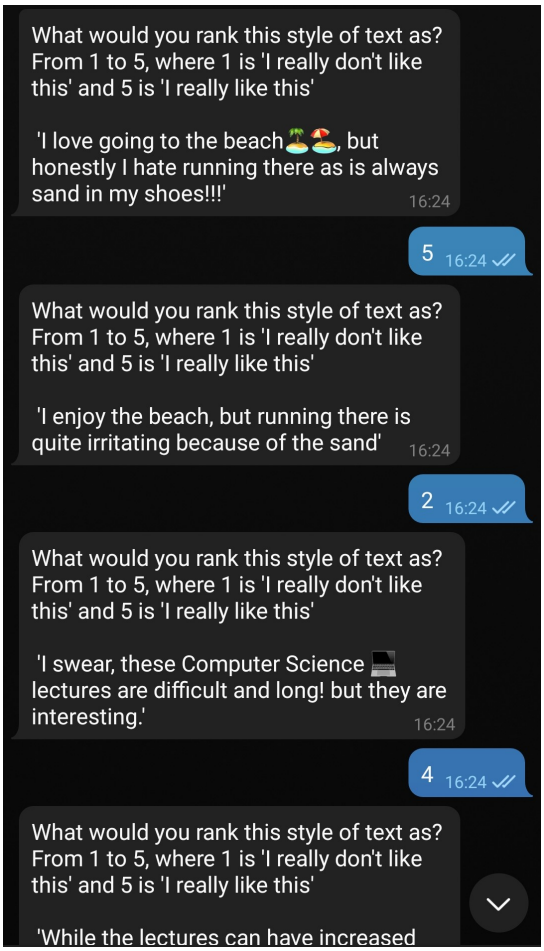

Figure 2: Example of the pre-task questions, designed to estimate the user's preferred conversation style

analyzing the total satisfaction score in a style alignment environment, the satisfaction score have a $M = 4$ and $SD = 1, N = 25$. In the environment where the style was not aligned, the satisfaction score was lower $M = 3.64$ and $SD = 1.25, N = 25$ as presented in Figure 4.

4.2 User engagement

Following the approach suggested by the works of [12], the method to obtain a total engagement score per user was as such: The scores of the three questions which were under the category of perceived usability were reversed, as they portray the negative engagement experience. As before, the participants who had no preference were excluded from this analysis. 12 people out of the 25 reported an increase of total engagement, when the

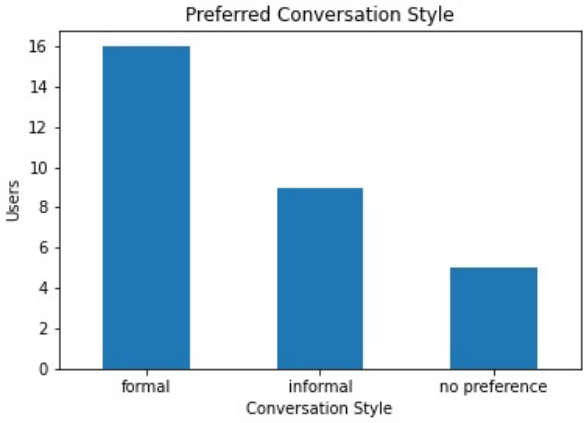

Figure 3: Number of users selecting each style

Figure 5: Averaged total engagement scores, with and without style alignment

Figure 4: Averaged satisfaction scores, with and without style alignment

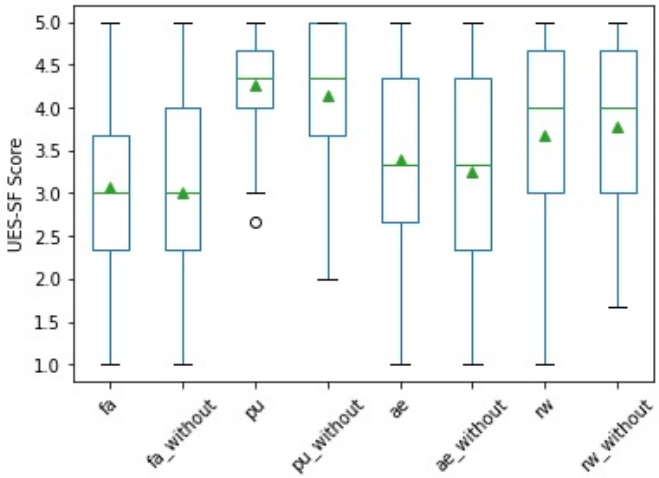

Figure 6: Engagement score per category, with style alignment and without style alignment

conversation style was aligned. For the participants who improved engagement, the improvement can be shown as $M = 0.583, SD = 0.55, N = 12$. When comparing the engagement score in the environment with style alignment and without, the style alignment can be described as $M = 3.6, SD = 0.82$. In the environment without conversation alignment, the results are almost identical $M = 3.55, SD = 0.93$ as can be seen in Figure 5. An analysis was also performed, on a UES-SF category basis, and the score of every category was averaged in both environments. The results showed minor differences and can be seen in Figure 6.

4.3 Participants With No Preference

Some participants indicated no preference towards any style and had equal scoring for each pair of pre-task statements. Their scoring for satisfaction can be seen in Table 2. The results of this group are inconclusive. No significant difference of ratings between the two alignment environments was evident.

5 Discussion

The experiment contained a small sample size which prevents any significant statistical findings. The results have no support for **Hypothesis 1** and produced near-identical engagement scores with or without style alignment, both in categorical and total scores. These results are likely to be related to the repetitive feedback provided to users in the informal style and to the lack of feedback in the formal style. which might be prejudicial to the natural conversation aspect, as it would resemble a conversation with a robot saying the same text in reply, as opposed to a conversation with a human. In the informal conversation style, a reply was chosen randomly from a list of three possible replies after each incoming message, without any consideration to the user's message. The formal environment provided no feedback to the user's message.

The same cannot be said about satisfaction, which

worker	UES-SF Informal	UES-SF Formal	Satisfaction Informal	Satisfaction Formal
0	3.92	3.58	4	4
1	3.92	3.83	4	5
2	3.17	3.17	3	3
3	4.08	2.25	5	2
4	2.92	3.17	3	3

Table 2: Scoring of participants without a preference towards a conversation style. UES-SF scores are total.

results show to support **Hypothesis 2**, as the satisfaction was higher for 36% of the participants. The most interesting aspect of Figure 4 is the density of higher satisfaction ratings in the style alignment environment.

Interestingly, a large majority of participants have preferred the formal conversation style, this result is somewhat counterintuitive. This finding is contrary to previous studies which have suggested to incorporate informal conversation elements in the CA [23, 10]. Our results align with previous research on the appropriate conversation style usage in a CA [11, 6]. As the experiment participants had no previous familiarity with the Dandelion bot, it could serve as the reason for the large percentage of participants who preferred formal style. These findings highlight the importance of adaptive chatbots, which would allow for dynamic style alignment, as users get more familiar with a specific chatbot, and after a certain number of interactions might prefer informal style.

5.1 Verbal Feedback

We have collected text feedback from users after the evaluation, to analyze further insights that could benefit the implications for future designers of CAs. It is worth mentioning that during the study, the participants were not made aware of the current conversation style or which one they received first.

Some users reported the lack of feedback from the formal conversation to be rude, and that could have influenced the engagement reported by the users.

> "The conversation felt robotic. This made the conversation feel like I was talking to a bot which doesn't care about what I think."

A few users mentioned that they have enjoyed the feedback provided by the informal bot. If the user's answers were accepted and matched the required answer format, the bot responded from pre-written replies.

> "I really liked the messages sent by the bot. I knew my messages were accepted and it felt more like a conversation than a survey."

There were also some comments regarding the rapid response time of the bot. Since the bot responded with no delay between messages, some users missed the feedback provided for their answers.

> "The bot sent messages so fast, only in middle of survey 1 I saw he replied to my messages."

This finding is consistent with that of Gnewuch et al. who studied the effect of intentional and dynamic delay in conversational agents [25]. They found that delay in bot response lead to better satisfaction rates, as the conversation possessed more human traits and resembled a human-to-human conversation, rather than an automatic, scripted conversation. However, the effects of delay were tested in the context of customer service. There might be a different outcome if the expectation of the users is for an instant reply from the CA. The effects of conversation delays in different contexts are beyond the scope of this paper.

5.2 Limitations

This study suggests a method to estimate to user's preferred conversation style, by asking a pre-task survey, containing four statements. Grave efforts were made to keep each pair of statements identical in meaning, but dissimilar in the style of writing. No pretest for reliability was performed in this study due to time constraints. The possibility of other attributes rather than style influencing the ranking cannot be ruled out.

A technical constraint was posed by the Dandelion system. In its original implementation, Dandelion did not support Neuro-linguistic programming (NLP). Therefor, when Dandelion received answers to questions asked during a survey, it had no technical ability to gather an intent and understand what the answer is about. Such constraint limited the ability to respond with an appropriate response to the user's answers. We used pre-written responses which were picked at random with equal probability.

Due to budget constraints, relatively small sample size was selected for this research (N=30). As with any experiment which revolved unpredictable variables, a larger sample size could mitigate any effects caused by outliers.

5.3 Design Implications

Our study found that using a dynamically aligned conversation style matched to the user's preference can have benefits such as an increase in satisfaction levels. While our evidence does not show that conversation style

alignment can increase engagement, we believe these results might have been influenced by other CA design elements that were discussed in Section 5.2. By adopting a conversation style that matches the user's preferred style, CA creators may experience recurrent interaction from the users [6].

If engagement marks an important role in a CA design, creators need to contemplate on other contributing factors, such as intended message delay [25] and integrating NLP to gather intent and provide proper conversational social cues [26]. Given these factors, more research is needed combined with conversational style alignment to measure the potential difference in engagement.

5.4 Future Work

Implicit Detection of Preferred Style

This research suggests a method to identify the user's preferred conversation style through a pre-task survey. While it does have the benefit of potentially identifying the user's preferred conversation style, it does require more time from the user. Another possible option is to perform the pre-task not before every task, but perhaps, every other task, as a user's preferred conversation style might not change that often. We would also like to mention the possibility of implicitly detecting the user's preferred conversation style, without explicitly asking for it.

An exploratory study experimented with different conversation styles in the context of conversational agents [27]. Their results reside with the theory concept laid by Tannen [9], where users would prefer talking with another entity that has a similar conversation style as their own. These results open up the possibility of analyzing the user's previously stored answers to best try to estimate the preferred style. This approach, while being transparent from the user, does require an extensive amount of saved users' responses, which might not always be available. More research is needed on implicitly detecting the user's preferred conversation style.

Context Influenced Conversation Style

The conversation style preferred by the user can also be influenced by the context of the conversation [19]. A survey asking a user to rate and provide feedback on a university course might foster different emotions to a survey asking the user about his personal health records, or personal well-being. More research is needed, to better understand the specific desires of a user's preferred conversation style regarding different conversation topics or domains.

6 Responsible Research

6.1 Reproducible Research

The increasing number of experiments that cannot be reproduced or verified can obstruct progress in the field of Computer Science. Research that deals with dynamic and inconsistent experiment variables such as human beings can challenge a reproduction attempt made at a later stage. However, it is the authors' responsibility to create the most capable environment for future reproduction trials. The first recommendation made to the scientists creating the experiment is to provide the source code, version, and data used in the experiment [28].

Therefore, a GitLab [1] repository containing the questions used in this experiment and the pre-task survey, was made available. The Dandelion system is at the time of writing, unavailable to the public. However this experiment could be recreated on many different chat platforms that support bots. No special features of Telegram bots were used in this study. The complete rankings and preferences of the participants were also made available. Due to privacy concerns, the Prolific id of the participants was removed from the repository data.

6.2 Ethical Research

The experiment conducted included human participants, thus strict guidelines were followed which were set by the TU Delft Human Research Committee. These guidelines were followed to assure no harm was done to any participant. As the participants were provided by the Prolific platform [22], no personally identifiable details were exposed to the research team. Before enrolling in the experiment, every participant had to read the description of the experiment, where the details and goals of the experiment were made available. It was stated that in case the participant would want to retract his participation and have any input provided by him deleted, a cancellation command was provided and implemented in the Dandelion system. The surveys filled by the participants contained general well-being questions, which might be considered as personal information. The answers to these questions were not examined in any analysis and were removed from the system after the experiment was finalized.

Employing participants from online platforms can have substantial difficulties in determining a fair price for the participants' output. Not only do participants devote different effort and time from one another when participating in the experiment, but they also live in different countries, with different labor laws and minimum wage. Many studies have tried formulating the issue and offered techniques to calculate a "fair pricing" [29] for the workers' output and effort [30, 31]. The workers were paid according to the time it took them to complete the experiment design as mentioned in section 3. They were paid an hourly wage of £6.30 per hour.

7 Conclusion

The aim of the present research was to examine the influence of dynamically aligning the conversation

[1] https://gitlab.com/barlerer/cse3000-conversation-styles-research

style in a CA environment as a technique to increase engagement and satisfaction. To identify the user's desirable conversation style, we have suggested a pre-task survey method, composing of ranking of four statements, where each pair is near-identical in meaning but differs in writing style.

The results of the experiment including 30 participants indicated that while dynamically aligning the conversation style has little to no influence on the perceived engagement, there is a noticeable difference in satisfaction. 36% of users reported higher satisfaction levels when the conversation style matched their pre-survey ranking. We have outlined relevant suggestions for future designers of conversational agents.

References

[1] Amon Rapp, Lorenzo Curti, and Arianna Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151, 03 2021. doi: 10.1016/j.ijhcs.2021.102630.

[2] Bayan Shawar and Eric Atwell. Chatbots: Are they really useful? *LDV Forum*, 22:29 – 49, 01 2007.

[3] S. Kiesler, J. Siegel, and T. McGuire. Social psychological aspects of computer-mediated communication. *Computer Supported Cooperative Work*, pages 657–682, 1984.

[4] Q. Liao, Muhammed Mas ud Hussain, P. Chandar, Matthew Davis, Y. Khazaeni, M. Crasso, Dakuo Wang, Michael J. Muller, N. Shami, and Werner Geyer. All work and no play? conversations with a question-and-answer chatbot in the wild. In *CHI 2018*, 2018.

[5] Deborah Tannen. New york jewish conversational style. 1981(30):133–150, 1981. doi: doi:10.1515/ijsl.1981.30.133.

[6] Christine Liebrecht, Lena Sander, and Charlotte van Hooijdonk. Too informal? how a chatbot's communication style affects brand attitude and quality of interaction. In *Conversations 2020: 4th international workshop on chatbot research*, 2020.

[7] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Chatterbox: Conversational interfaces for microtask crowdsourcing. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '19, pages 243 – 251, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450360210. doi: 10.1145/3320435.3320439.

[8] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1 – 12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300316.

[9] Deborah Tannen. *Conversational style: analyzing talk among friends*. Oxford University Press, 2005.

[10] Christine Liebrecht and Charlotte van Hooijdonk. Creating humanlike chatbots: What chatbot developers could learn from webcare employees in adopting a conversational human voice. In Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg, editors, *Chatbot Research and Design*, pages 51–64, Cham, 2020. Springer International Publishing. ISBN 978-3-030-39540-7.

[11] Anais Gretry, Csilla Horvath, Nina Belei, and Allard C.R. van Riel. "don't pretend to be my friend!" when an informal brand communication style backfires on social media. *Journal of Business Research*, 74:77–89, 2017. ISSN 0148-2963. doi: https://doi.org/10.1016/j.jbusres.2017.01.012.

[12] Heather L. O'Brien, Paul Cairns, and Mark Hall. A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies*, 112:28–39, 2018. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2018.01.004.

[13] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376403.

[14] Walter S. Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F. Allen, and Jeffrey P. Bigham. Chorus: A crowd-powered conversational assistant. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 151–162, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450322683. doi: 10.1145/2501988.2502057.

[15] Ting-Hao Huang, Walter Lasecki, and Jeffrey Bigham. Guardian: A crowd-powered spoken dialog system for web apis. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 3(1), Sep. 2015.

[16] Rucha Bapat, Pavel Kucherbaev, and Alessandro Bozzon. Effective crowdsourced generation of training data for chatbots natural language understanding. In Tommi Mikkonen, Ralf Klamma, and Juan Hernández, editors, *Web Engineering*,

pages 114–128, Cham, 2018. Springer International Publishing. ISBN 978-3-319-91662-0.

[17] Heather O'Brien and Paul Cairns. An empirical evaluation of the user engagement scale (ues) in online news environments. *Information Processing & Management*, 51(4):413–427, 2015. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2015.03.003.

[18] Allison W. Harrison and R. Kelly Rainer. A general measure of user computing satisfaction. *Computers in Human Behavior*, 12(1):79–92, 1996. ISSN 0747–5632. doi: https://doi.org/10.1016/0747-5632(95)00020-8.

[19] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Estimating conversational styles in conversational microtask crowdsourcing. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), May 2020. doi: 10.1145/3392837.

[20] Dandelion website. https://delftdandelion.com.

[21] Sihang Qui, Willem van der Maden, James Derek Lomas, and Ujwal Gadiraju. Context-Sensitive Assessments of Human Wellbeing. In *CHI'21 Workshop: The New Vulnerable*, March 2021.

[22] Prolific website. https://www.prolific.co.

[23] Jurek Kirakowski, Patrick O'Donnell, and Anthony Yiu. Establishing the hallmarks of a convincing chatbot-human dialogue. *Human-computer interaction*, pages 49–56, 2009.

[24] Justine Cassell and Timothy Bickmore. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modelling and User-Adapted Interaction*, 13, 04 2001. doi: 10.1023/A:1024026532471.

[25] Ulrich Gnewuch, Stefan Morana, Marc Adam, and Alexander Maedche. Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction. 06 2018.

[26] Ryan M. Schuetzler, G. Mark Grimes, and Justin Scott Giboney. The impact of chatbot conversational skill on engagement and perceived humanness. *Journal of Management Information Systems*, 37(3):875–900, 2020. doi: 10.1080/07421222.2020.1790204.

[27] Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory Bennett. An exploratory study toward the preferred conversational style for compatible virtual agents. In *Intelligent Virtual Agents*, pages 40–50. Springer, January 2017.

[28] Reproducible research. *Computing in Science & Engineering*, 12(05):8–13, sep 2010. ISSN 1558-366X. doi: 10.1109/MCSE.2010.113.

[29] Ria Mae Borromeo, Thomas Laurent, Motomichi Toyama, and Sihem Amer-Yahia. Fairness and Transparency in Crowdsourcing. In *International Conference on Extending Database Technology (EDBT)*, Venice, Italy, March 2017. doi: 10.5441/002/edbt.2017.46.

[30] Chenxi Qiu, Anna Squicciarini, and Benjamin Hanrahan. Incentivizing distributive fairness for crowdsourcing workers. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 404–412, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

[31] Zehong Hu and Jie Zhang. Optimal posted-price mechanism in microtask crowdsourcing. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 228–234, 2017. doi: 10.24963/ijcai.2017/33.